LA INFERENCIA ESTADÍSTICA

FISHER

Probablemente sí, probablemente no



NATIONAL GEOGRAPHIC

RONALD FISHER fue el científico que dotó a la estadística de las herramientas que le han permitido alcanzar la enorme dimensión con la que cuenta en la actualidad. La inferencia estadística, su mayor contribución, introdujo una pieza novedosa, relacionada con la probabilidad, que tuvo el poder de insuflar el oxígeno necesario para que esta materia, hasta ese momento una simple herramienta al servicio de otras disciplinas, se convirtiera en una ciencia por derecho propio. A este matemático y biólogo británico se debe el empleo del método estadístico en el diseño de experimentos científicos. Sus investigaciones también se adentran en la genética y en la teoría evolutiva moderna, en un contexto marcado por la eugenesia, muy en boga en la primera mitad del siglo XX y de la que era un ardiente defensor.

LA INFERENCIA ESTADÍSTICA

FISHER

Probablemente sí, probablemente no

NATIONAL GEOGRAPHIC

CARLOS M. MADRID CASADO es profesor de Estadística en la Universidad Complutense de Madrid e investigador asociado de la Fundación Gustavo Bueno para temas de historia y filosofía de las ciencias.

Realización: EDITEC

Diseño cubierta: Llorenç Martí

Diseño interior: Luz de la Mora

Infografías: Joan Pejoan

Fotografías: American Philosophical Society: 119; Archivo RBA: 59, 111, 138; Archivos Federales de Alemania: 80; Departamento de Estadística/Universidad de Berkeley, California: 150; Elliott & Fry/National Portrait Gallery: 45; Galton.org: 37, 61a, 119; Gonville & Caius College, Cambridge: 79a; Instituto Internacional de Estadística: 157b; Konrad Jacobs: 130; LSE Library: 141; Museo Americano de Historia Natural: 121; National Portrait Gallery, Londres: 61bi, 135; Rothamsted Research: 79b; John Snow: 27; Universidad de Adelaida, Australia: 33, 157ai, 157ad; University College, Londres/Cold Spring Harbor Laboratory: 61bd.

ISBN: 978-84-473-7776-3
Depósito legal: B-20896-2016

Impreso y encuadernado en Rodesa, Villatuerta (Navarra)

Impreso en España - *Printed in Spain*

Sumario

Introducción

En los manuales la estadística suele definirse como la ciencia que estudia la recogida, organización e interpretación de datos. Pero en esta definición brilla por su ausencia un componente esencial: el trabajo estadístico se realiza empleando el lenguaje de la probabilidad. La estadística aborda el estudio probabilístico de la incertidumbre, sea cual sea su fuente. Así, por ejemplo, la inferencia estadística se ocupa de evaluar y juzgar las discrepancias observadas entre la tozuda realidad y lo prescrito por el modelo teórico, haciendo uso indispensable del cálculo de probabilidades. Pero, ¿quién fue el responsable de la inyección conceptual y probabilística que experimentó la estadística decimonónica a principios del siglo XX?

La estadística tiene muchos próceres: Karl Pearson, Jerzy Neyman o Abraham Wald son algunos de ellos. Pero solo tiene un genio: Ronald Aylmer Fisher. Un gran número de las técnicas estadísticas hoy habituales tiene su origen en la obra de sir Ronald, aunque la mayoría de libros de texto omitan esta deuda. La lectura de los artículos y los libros de Fisher, donde la discusión lógica o filosófica siempre encuentra espacio entre el desarrollo matemático, resulta ilustradora, sorprendente y, a menudo, comporta la exasperación del lector, por cuanto el estadístico británico hacía gala de un estilo mordaz e insolente para con muchos de sus colegas, sin escatimar insultos. Pero acercarse a la figura

de Fisher supone asistir a la fábrica de la estadística matemática moderna.

Las aportaciones más descollantes de nuestro personaje emergieron en un trasfondo histórico de lo más enrevesado, conformando un mosaico de conceptos científicos e ideas filosóficas. Fisher bebió de las fuentes de la estadística a través de tres ciencias por completo diferentes: por medio de la astronomía conoció las contribuciones de Gauss y Laplace; la física de gases le enseñó las aplicaciones desarrolladas por Quetelet y Maxwell, y, finalmente, la biología evolutiva le abrió las puertas de las principales novedades estadísticas de finales del siglo XIX, que llevaban la firma de Francis Galton y Karl Pearson.

Se antoja imposible calibrar la verdadera talla de Fisher sin compararlo con ese titán llamado Karl Pearson. En su búsqueda de una teoría matemática de la evolución, Pearson ideó algunos de los métodos estadísticos hoy clásicos. Sin embargo, fue demasiado lento a la hora de reconocer el talento de Fisher, adoptando una cerrazón recalcitrante ante las rectificaciones que el joven y astuto investigador introducía a su propio trabajo. Pearson pagó caro su error, porque los artículos de juventud de Fisher enseñaron nuevos horizontes, ensanchando el mundo estadístico conocido y preparando la eclosión de la inferencia estadística.

Fisher tenía diecinueve años cuando ingresó en la Universidad de Cambridge y veintinueve cuando, en 1919, aceptó un puesto como estadístico en la Estación Agrícola Experimental de Rothamsted. Allí, rodeado de patatas, fertilizantes y ratones, cimentó gran parte del éxito y la fama de su carrera investigadora. Durante los años veinte, Fisher recogió el testigo de la oleada de estadísticos crecida en torno a Karl Pearson, consolidando el estatuto científico de la estadística al cohesionar sus fundamentos matemáticos. El estadístico inglés la dotó de una serie de conceptos y métodos característicos. El vocabulario técnico que redefinió o acuñó para la ocasión es solo la punta del iceberg: población, muestra, parámetro, estadístico, varianza, verosimilitud, prueba de significación, aleatorización...

Fisher fue el arquitecto que, simultáneamente, puso los pilares de la teoría de la estimación y de la teoría de los test estadísti-

cos. Mientras que la primera se centra en determinar un estimador apropiado para cada parámetro desconocido, así como de comparar las propiedades de los candidatos, la segunda se preocupa de someter hipótesis que establezcan valores concretos del parámetro al dictado de la experiencia. Cuando un astrónomo realiza repetidas mediciones de la posición de una estrella y quiere predecir su posición real, emplea la teoría de la estimación. Cuando dos astrónomos mantienen valores diferentes para la posición de la estrella y deciden realizar una observación conjunta para salir de dudas, emplean la teoría de los test estadísticos. Pero hay más. Fisher es el creador de lo que los estadísticos denominan «diseño de experimentos», es decir, del uso de la estadística en el momento de planear cualquier experimento.

Todo este espléndido bagaje se dio a conocer en el libro *Métodos estadísticos para investigadores*, publicado en 1925, cuyo impacto fue tremendo. No tanto por las ventas que cosechó, sino por la cantidad de investigaciones que motivó, y no solo entre estadísticos y matemáticos, sino principalmente entre ingenieros agrónomos, biólogos, químicos y científicos en general. La estadística había llegado para quedarse.

Esta panorámica no estaría completa si no se mencionase que la genética fue la otra disciplina que, junto con la estadística, acaparó los pensamientos de Fisher de por vida. Nuestro autor es uno de los fundadores de la genética de poblaciones, la ciencia que permitió reconciliar a Darwin con Mendel, es decir, la selección natural de las especies con las leyes de la herencia, asentando de esta manera la teoría sintética de la evolución o neodarwinismo. No obstante, el interés de nuestro personaje por el tema venía suscitado por la eugenesia, una inquietante doctrina —colindante con el racismo— que marcó la primera mitad del siglo pasado, pero que para Fisher hizo de gozne entre la estadística y el evolucionismo.

A lo largo de este libro también nos acercaremos a las numerosas controversias científicas y filosóficas en que se sumergió Fisher, muchas de las cuales aún perduran, y que son una prueba más de la vitalidad de la estadística. La teoría estadística clásica, tal como hoy la conocemos (conteniendo la estimación, el con-

traste de hipótesis, el diseño de experimentos y el muestreo), es fruto de dos hombres: Ronald Aylmer Fisher y Jerzy Neyman, cuyas contribuciones muchas veces aparecieron en paralelo, complementándose pero también contradiciéndose. A ninguno de los dos estadísticos le gustó nunca ver asociado su nombre al del rival, pese a que al comienzo mantuvieron una relación amistosa. El rabioso antagonismo entre ambos no terminó hasta la muerte de Fisher, porque para este las aportaciones de Neyman no hacían sino corroer las suyas propias.

El estadístico británico reflexionó profundamente sobre el papel que corresponde a la inferencia estadística en el método científico, entrando con ello en polémica con la mayoría de sus colegas. Uno de los problemas favoritos de los filósofos, de Aristóteles a Hume, se convirtió en idea fija del pensamiento fisheriano. Nos referimos, claro está, al problema secular de la inducción, que él concatenó con la probabilidad y la estadística. Las inferencias inductivas establecían, por así decir, conclusiones probabilísticas.

Supongamos por un instante que somos médicos y nos planteamos, a propósito de un paciente, la hipótesis de si padece tuberculosis. De cara a examinar la validez de esta hipótesis, le hacemos una prueba rutinaria con rayos X que da negativa. Obviamente, este resultado no es concluyente, porque toda prueba médica puede fallar, presentando lo que suele denominarse un «falso negativo» (de la misma manera que a veces se obtienen «falsos positivos»). Nos encontramos, pues, ante un genuino test estadístico. En esta situación podemos formularnos tres preguntas distintas:

1. A partir del dato, ¿qué debemos creer y en qué grado? ¿Cuál es la probabilidad de que el paciente tenga tuberculosis sabiendo que ha dado negativo en el test?

2. ¿Qué información aporta el dato sobre la verosimilitud de la hipótesis? ¿Podemos inferir que no presenta la enfermedad?

3. Dado el dato, ¿qué debemos hacer? ¿Aceptamos o rechazamos la hipótesis de que tiene tuberculosis?

Mientras que la primera pregunta se centra en la *creencia*, la segunda lo hace en la *evidencia* y la tercera en la *decisión*. Como tendremos ocasión de explicar, Fisher intentó responder al segundo enigma. Los estadísticos bayesianos contestan, por su parte, al primero, y los estadísticos que siguen las enseñanzas de Neyman lo hacen al tercero. Bayesianos y frecuentistas —incluyendo bajo este rótulo tanto a los partidarios de Fisher como de Neyman— aglutinan los dos polos que roturan el campo de la estadística.

Es un hecho que la aportación de Fisher cambió el paradigma científico de la época; pero no es fácil discurrir el modo en el cual la estadística se convirtió por su mano en una ciencia *per se*, en una disciplina autónoma, partiendo de ser un apéndice de otras disciplinas como la astronomía, la sociología o la biología. La naturaleza de la estadística, que engloba contenidos y aplicaciones de lo más diverso, es sumamente problemática y para nada resulta sencillo determinar cuál es el nexo que dota de unidad a su campo, más allá de un ramillete de herramientas matemáticas.

La convergencia de varias disciplinas naturales y sociales posibilitó la configuración de la estadística y, al mismo tiempo, aunque resulte paradójico, su emancipación respecto de ellas. Desde los juegos de azar, las leyes estadísticas —cuya regularidad se revela a la escala del colectivo, no del individuo— se radiaron a la astronomía y la geodesia, la sociología, la biología, la agricultura, la industria, etcétera. Las monedas, los dados, las barajas y las urnas son el modelo que utilizamos para razonar estadísticamente sobre los astros, las personas, los genes, las cosechas o la producción de coches. Para los antiguos, la probabilidad y la estadística aparecían en la observación de la naturaleza. Desde Fisher lo hacen preferiblemente en el muestreo, cuando se extrae una muestra aleatoria de una población, aunque esta última no sea más que un producto de la imaginación del estadístico.

Ronald Aylmer Fisher hizo de la estadística una ciencia a medio camino entre la matemática y la experiencia, donde la confrontación con problemas tangibles estimula su crecimiento tanto o más que los problemas teóricos. Son los materiales demográficos, económicos o sanitarios los que constituyen esta ciencia y le otorgan su preeminencia actual. Sin su estigma se reduciría a una

disciplina marginal, teorética. La estadística se entreteje con una pléyade de ciencias experimentales, proyectando luz sobre sus campos y funcionando, muchas veces, como una suerte de geometría de las inferencias. Solo así se comprende cómo ha conquistado casi todos los espacios a lo largo del siglo XX. Su irrupción se inserta dentro de la gran revolución tecnológica del siglo pasado. Es un patrón de objetividad y estandarización que se aplica en las mediciones oficiales, los procesos de fabricación o las investigaciones farmacéuticas. Sirva como ejemplo que la noción de una población como una cifra exacta apenas tuvo sentido hasta que no hubo instituciones estadísticas encargadas de definir lo que significa y de establecer con precisión cómo estimar el número de habitantes, trabajadores o votantes de un país. La estadística ha generado un mundo que se ha ido haciendo numérico hasta el último de sus rincones.

Y la chispa de este fuego que hoy nos calienta la encendió, desde luego, nuestro protagonista. Un científico excepcional, en su inteligencia y en su arrogancia. Nadie como él ahondó tanto en los fundamentos de la estadística. Su obra es la columna vertebral de la ciencia que hoy conocemos. Ahora, cojan aire y prepárense para bucear en el océano de la ciencia estadística.

1890 Ronald Aylmer Fisher nace el 17 de febrero en una localidad del extrarradio de Londres.

1909 Ingresa en la Universidad de Cambridge, donde estudia matemáticas, astronomía, mecánica estadística, teoría cuántica y biología.

1915 Fisher se anota su primer gran tanto al deducir la distribución del coeficiente de correlación en el muestreo. La demostración se publica en *Biometrika*, la revista editada por Karl Pearson.

1917 La sintonía entre Fisher y Pearson comienza a resquebrajarse como consecuencia de las ásperas críticas que se dirigen.

1919 Fisher ingresa en la Estación Agrícola Experimental de Rothamsted.

1922 Plantea los conceptos centrales de la inferencia estadística en su artículo «Sobre los fundamentos matemáticos de la estadística teórica».

1925 Publica *Métodos estadísticos para investigadores*, uno de los libros que más ha hecho por la implantación y difusión de la estadística entre científicos e ingenieros.

1930 Aparece la monografía *La teoría genética de la selección natural*, donde demuestra que la herencia mendeliana es compatible con el darwinismo.

1933 Tras el retiro de Karl Pearson, Fisher se hace con el control de la mitad del departamento que lideraba en el University College de Londres: la cátedra de Eugenesia. La cátedra de Estadística pasa a manos del hijo, Egon Pearson.

1935 Se publica *El diseño de experimentos*, libro de cabecera para los científicos que querían sacar el máximo partido a sus experimentos empleando herramientas estadísticas. Se inicia la polémica con Jerzy Neyman y Egon Pearson a propósito de las pruebas de significación y los contrastes de hipótesis.

1943 Regresa a Cambridge para ocupar la cátedra de Genética.

1955 Los rescoldos de la disputa mantenida con Neyman y Pearson se reavivan intensamente con motivo del artículo incendiario que Fisher presenta en la Real Sociedad de Estadística sobre la inferencia inductiva.

1958 Fisher polemiza sobre la relación entre el hábito de fumar y el cáncer de pulmón, negando que se haya demostrado su asociación.

1962 Muere, como consecuencia de un cáncer de colon, el 29 de julio en Adelaida (Australia), donde pasó sus últimos años de vida como investigador emérito.

CAPÍTULO 1

La estadística antes de Fisher

A finales del siglo XIX los métodos estadísticos se encontraban desperdigados por varios campos bastante distanciados. La astronomía custodiaba las aportaciones de Gauss y Laplace relativas al método de mínimos cuadrados, la ley del error y el cálculo de probabilidades. La curva normal era de uso común en la sociología y en la física de gases, gracias a la semejanza entre las moléculas de un gas y los ciudadanos de un país. Pero sería dentro del perímetro de la biología evolutiva donde aparecerían las principales novedades estadísticas del siglo.

Ronald Aylmer Fisher nació el 17 de febrero de 1890 en East Finchley (Londres). Sus padres, tras el nacimiento de sus dos primeros hijos (Geoffrey y Evelyn), decidieron llamar a su tercer hijo Alan, pero su temprana muerte les hizo adoptar una llamativa superstición: todos sus hijos sin excepción llevarían una «y» en el nombre, incluyendo el más joven de los siete que tuvieron, Ronald Aylmer. Desde muy pequeño Ronald demostró tener un talento especial para las matemáticas. Con seis años, su madre comenzó a leerle un libro divulgativo de astronomía, que despertó en él un interés que no le abandonó en la infancia ni en la adolescencia. Sin embargo, desde los días de la escuela, su vista mostró ser muy pobre: padecía una miopía extrema, de manera que los médicos le prohibieron estudiar con luz eléctrica, artificial. Durante las tardes, los profesores particulares le enseñaban sin lápiz ni papel, lo que le permitió desarrollar una habilidad excepcional para resolver problemas matemáticos de cabeza, basándose en intuiciones geométricas pero omitiendo los detalles (una costumbre que le acompañó toda la vida).

Cuando tenía catorce años, su madre murió de un ataque agudo de peritonitis y, poco después, su padre perdió toda su fortuna. Por suerte, Fisher ganó una beca para financiarse la universidad. En Cambridge, donde ingresó en 1909, estudió matemáticas y astronomía, aunque también se interesó por la biología. Tras

graduarse, completó sus estudios dentro del campo de la «teoría de errores», una teoría matemática de gran utilidad en astronomía y que constituyó, junto con la teoría de gases, su primer contacto con la estadística. Puede parecer paradójico que el creador de la estadística matemática moderna conociese la disciplina que contribuyó a revolucionar por medio de la astronomía, como si los astros guardasen el secreto de las encuestas o las elecciones. Para poder explicar este hecho, y con él la magnitud de la obra de Fisher, es obligado volver la vista atrás, al siglo XIX, y rastrear el origen de los métodos estadísticos a través de varias disciplinas fronterizas.

Generalmente se admite que la estadística se divide en dos ramas bien diferenciadas pero interconectadas. Por un lado, la estadística descriptiva, que se encarga del análisis exploratorio de datos; por otro, la estadística inferencial (o inferencia estadística), encaminada a hacer predicciones en situaciones de incertidumbre. El germen de la estadística inferencial se encuentra en los juegos de azar y en la astronomía, aunque el conjunto de conceptos que se desarrollaron tardó en circular al ámbito social en que brotó la estadística descriptiva. Esta primera fase abarca, aproximadamente, desde 1650 a 1850. Finalmente, en una segunda fase, coincidiendo con la segunda mitad del siglo XIX, las herramientas estadísticas conocieron una nueva circulación: de la astronomía y la sociología a la biología. Pero comencemos por el principio.

DE LAPLACE A LA SOCIALIZACIÓN DE LA ESTADÍSTICA

Podemos imaginar la ciencia estadística como un río formado por la confluencia de dos afluentes que discurrían independientes. Por una parte, el cálculo de probabilidades, que es la base de la inferencia estadística. Por otra, «la ciencia del Estado», de donde deriva precisamente el nombre «estadística», y que tiene más que ver con la estadística descriptiva.

El cálculo de probabilidades surgió, pese a las aportaciones pioneras de Cardano, Galileo y algunos escolásticos, al calor de

los juegos de azar ya avanzado el siglo XVII. Cartas, dados, monedas y urnas funcionaron como paradigmas de la naciente «geometría del azar», según atestigua la correspondencia que a partir de 1654 entablaron un austero jansenista y un abogado amante de las matemáticas, Blaise Pascal y Pierre de Fermat, a propósito de los acertijos propuestos por Antoine Gombaud, caballero de Méré y jugador empedernido. El concepto de probabilidad —que como vocablo ya puede encontrarse en Cicerón— se les escapó a los griegos por carecer de una aritmética simbólica adecuada, así como de dados simétricos (los posibles resultados de su astrágalo no eran equiprobables), lo que les impidió postular la regla de Laplace —que ya se encuentra en Jakob Bernoulli (1654-1705) o Abraham de Moivre (1667-1754)— como axioma, y cuyo enunciado es el siguiente: «La probabilidad de un suceso es igual al número de casos favorables dividido por el número de todos los casos posibles». Ahora bien, conviene aclarar que el concepto de probabilidad tampoco aparece en las cartas que cruzaron Pascal y Fermat, y hay que esperar al *Ars Conjectandi* de Bernoulli, publicado póstumamente en 1713, para encontrar una discusión explícita de la noción.

En esta obra, Bernoulli partió de los problemas que había abordado Christiaan Huygens en su libro *De Ratiociniis in Ludo Aleae* (1657), aplicó la combinatoria a su resolución y, lo que es más importante en relación con la estadística, presentó el «teorema áureo» (una versión de la ley de estabilidad de las frecuencias) y discutió por vez primera el problema de la probabilidad inversa: ¿cuántas observaciones hacen falta para estimar una probabilidad a partir de la frecuencia? El matemático suizo fue pionero en plantearse la posibilidad de inferir la probabilidad de un suceso *a posteriori* (a partir de la experiencia) cuando no puede deducirse *a priori* (antes de la experiencia, mediante razonamientos lógicos o psicológicos).

A caballo entre los siglos XVIII y XIX, Pierre-Simon de Laplace (1749-1827) completó estos avances, fusionando el cálculo algebraico de probabilidades con el análisis matemático en su obra *Teoría analítica de las probabilidades* (1812). Si antes de él, con contadas excepciones, el cálculo de probabilidades se servía del álgebra, a partir de él lo haría básicamente del análisis, por medio

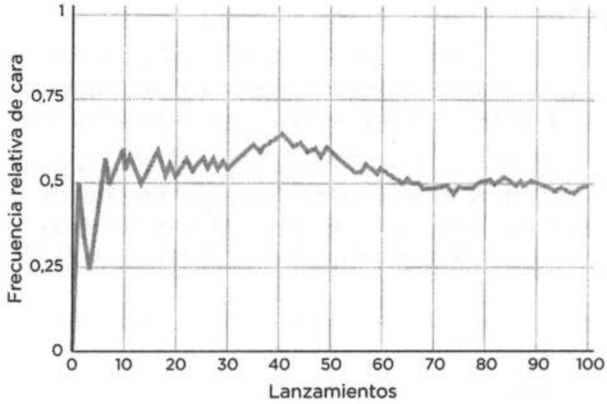
EL TEOREMA ÁUREO DE BERNOULLI

Este teorema, conocido hoy simplemente como teorema de Bernoulli, afirma que la frecuencia relativa de un suceso tiende a aproximarse a un número fijo —la probabilidad del suceso— conforme aumenta el número de repeticiones del experimento aleatorio. Formalmente: dados un suceso A, su probabilidad p de ocurrencia y n pruebas independientes para determinar la ocurrencia o no ocurrencia de A; si f es el número de veces que se presenta A en los n ensayos y ε es un número positivo cualquiera, la probabilidad de que la frecuencia relativa f/n discrepe de p en más de ε (en valor absoluto) tiende a cero al tender n a infinito. Es decir:

$$\lim_{n\to\infty} P\left(\left|\frac{f}{n}-p\right|>\varepsilon\right)=0.$$

Recíprocamente, la probabilidad de que la frecuencia relativa se estabilice a largo plazo tiende a 1 (lo cual no quiere decir que, eventualmente, no pueda haber desviaciones, esto es, rachas contrarias, «cisnes negros»). Así, por ejemplo, la frecuencia relativa con que sale cara al lanzar al aire una moneda legal se acerca a 0,5 (su probabilidad) cuando la lanzamos un número suficiente de veces. En la época, el conde de Buffon lanzó 4 040 veces una moneda y obtuvo 2048 caras, es decir, el 50,69 % de las veces. Este teorema, por tanto, formalizaba la ley del azar o ley de estabilidad de la frecuencia: hay —por decirlo con un término debido a Bernoulli— «certeza moral» (probabilidad de 0,999) de que a la larga la frecuencia relativa de un suceso no se desvía significativamente de su probabilidad (véase la figura). Era la «ley de los grandes números» —empleando la expresión acuñada en el siglo XIX por Siméon Denis Poisson (1781-1840)— en su forma más sencilla. En efecto, mientras que el teorema de Bernoulli nos asegura que la frecuencia relativa con que sale cara al tirar una misma moneda sucesivas veces tiende a estabilizarse, la ley de los grandes números nos asegura que la frecuencia relativa con que se obtiene cara al lanzar sucesivas monedas también se estabiliza, aunque cada moneda tenga una probabilidad de cara

de las funciones generatrices. Laplace definió con rigor el concepto de probabilidad y discutió ampliamente el problema de la probabilidad inversa, redescubriendo el teorema de Bayes (solo llamado así por Augustus de Morgan muchos años después, que vindicó la prioridad de su compatriota). Además, sentó las bases de la inferencia estadística bayesiana, que empleó para predecir tasas de matrimonios y proporciones de nacimientos según



Frecuencia relativa de que salga cara tras 100 lanzamientos de una moneda.

distinta. P.L. Chebyshev y la escuela rusa continuarían el estudio de las leyes de los grandes números, que generalizan el teorema áureo. Para Bernoulli el teorema posibilitaba calcular empíricamente las probabilidades desconocidas. Permitía definir la probabilidad de una forma objetiva, invirtiendo el teorema. En efecto, si la frecuencia se aproxima a la probabilidad según crece el número de observaciones, ¿por qué no definir la probabilidad a partir de la frecuencia? Mediante el recurso a la inducción parecía factible definir la probabilidad como el límite de la frecuencia, y no ya hacerlo de una forma meramente lógica o subjetiva (como un grado de creencia). No obstante, el matemático francés afincado en Inglaterra —por su irredento calvinismo, era hugonote— Abraham de Moivre, famoso por su tratado *La doctrina del azar* (1718), defendía que la regularidad estadística que postulaba el teorema áureo necesitaba obligatoriamente del concurso de Dios para funcionar. Fisher, como tendremos ocasión de explicar, heredó esta crisis abierta en la interpretación de la probabilidad.

el sexo. Y utilizó la teoría de probabilidades en la resolución de múltiples problemas de la mecánica celeste: por ejemplo, para examinar la distribución de las órbitas de los cometas como si se tratara de una serie de cuerpos proyectados aleatoriamente en el espacio, como dados lanzados sobre una mesa. Sin embargo, la aplicación de mayor envergadura vino de la mano de la «teoría de errores» que en su día estudiara Fisher.

En el período que abarca de 1770 a 1820 se desarrollaron los métodos estadísticos básicos en conexión con la astronomía, ya que esta ciencia requería de un estudio cuidadoso del error. Se trataba de reducirlo al mínimo a la hora de estimar la posición de un planeta o una estrella a partir del conjunto de observaciones. Un astrónomo quiere determinar la posición real del astro tras haber realizado una serie de mediciones. Laplace interpretó que la posición real de la estrella funcionaba como causa de las posiciones observadas, dependiendo los errores del azar. En estos términos, mediante una utilización ingeniosa del teorema de Bayes,

EL TEOREMA DE BAYES

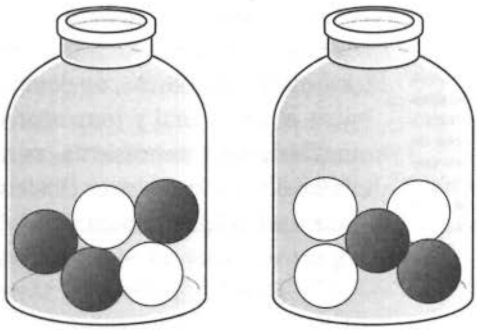
En una memoria de 1773 titulada «Sobre la probabilidad de las causas de los sucesos», Laplace se planteaba que las situaciones en las que interviene el azar son, generalmente, de dos tipos. En el primero, el azar aparece en los resultados. Por ejemplo: conocemos la composición de una urna en la que hay bolas blancas y negras, y nos planteamos cuál será el resultado de una extracción. A partir de las causas (la composición de la urna, que conocemos), calculamos la probabilidad de los resultados, de sacar blanca o negra. Hay, en cambio, un segundo tipo de situación en la que el azar no aparece en los resultados sino en las causas. Conocemos el resultado de la extracción (ha salido, pongamos por caso, una bola negra) y queremos calcular la composición de la urna, que nos es desconocida. A partir de los resultados (ha salido negra), determinamos la probabilidad de las causas, de cada posible composición de la urna. Pasamos, pues, de los efectos a las causas. Laplace enunció y demostró el teorema que descubrió el reverendo Thomas Bayes (1702-1761) y que se publicó en una memoria póstuma de 1763, pero que seguro desconocía (los matemáticos franceses no solían leer a los ingleses). Este teorema afirma que si $\{A_1, A_2, \ldots, A_n\}$ forman un conjunto de sucesos mutuamente excluyentes y exhaustivos, $P(A_i)$ son las probabilidades *a priori* de los sucesos y $P(B|A_i)$ son las verosimilitudes (la probabilidad de observar el efecto B supuesta la causa A_i), entonces la probabilidad *a posteriori* de cada suceso viene dada por:

$$P(A_i|B) = \frac{P(A_i)\cdot P(B|A_i)}{\sum_{k=1}^{n} P(A_k)\cdot P(B|A_k)}.$$

Lo que aquí nos interesa es explicar la idea latente tras la fórmula de Bayes que redescubrió Laplace, por cuanto fue uno de los caballos de batalla de

concluyó que existe una curva que representa la distribución del error en torno al valor real (figura 1, pág. siguiente). La curva es simétrica y decreciente a partir de ese valor central, en el sentido de que cuanto más nos alejamos de él menos probable es que cometamos tanto error al medir. En consecuencia, lo más probable es que el valor que elijamos como real (la media aritmética de los resultados) se encuentre en un entorno de ese valor central, donde la curva alcanza su máximo. Resolviendo una ecuación diferencial, Laplace llegó a que la curva de la distribución de los errores viene dada por una función de tipo exponencial.

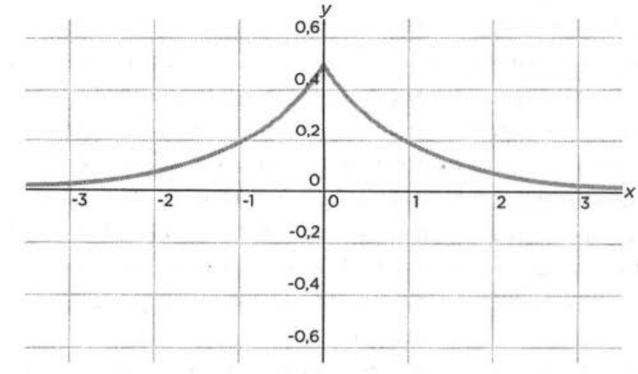
Fisher. Imaginemos una urna que puede tener dos composiciones diferentes: la primera contiene 2 bolas blancas y 3 bolas negras, y la segunda, 3 blancas y 2 negras, tal como muestra la figura. Se extrae una bola al azar y resulta ser negra, ¿qué composición de la urna es más probable? Intuitivamente, a la luz del color de la bola extraída, parece claro que la primera composición tiene que ser más probable que la segunda (dado que en esta última hay menos bolas negras). El teorema de Bayes no hace sino cuantificar numéricamente esta intuición. Las dos causas que han podido originar el suceso «sacar bola negra» son, precisamente, las dos posibles composiciones de la urna. Si se supone *a priori* que ambas composiciones son igualmente probables (0,5 para cada una de ellas), la utilización de la fórmula de Bayes lleva a que la probabilidad de la primera composición ha subido, tras la extracción de la bola negra, a 0,6, mientras que la probabilidad de la segunda composición ha bajado a 0,4. Las probabilidades *a priori* (0,5 y 0,5) han sido rectificadas *a posteriori* (0,6 y 0,4). Un resultado que parece incontrovertible, puesto que en la primera composición hay más bolas negras que en la segunda y, por lo tanto, cabe esperar una mayor probabilidad de que la bola haya sido extraída en esas condiciones. Para Laplace, al igual que para Bayes, este poderoso teorema posibilitaba aprender de la experiencia y, en el límite, legitimar la inducción.



Si hemos extraído una bola negra, el teorema de Bayes concluye que la probabilidad *a posteriori* de la composición de la izquierda es mayor que la de la derecha.

Ley de los errores según Laplace:
$$\phi(x)=\frac{e^{-|x|}}{2}.$$

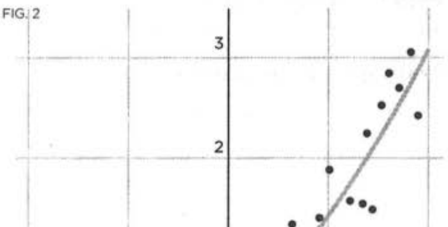
FIG. 1



El método de mínimos cuadrados sirve para ajustar sobre el conjunto de observaciones una trayectoria que minimice el error cuadrático.

Mientras que Laplace, a fin de combinar las observaciones sucesivas del astro en una trayectoria, buscaba minimizar la suma de los errores absolutos, es decir, de las diferencias en valor absoluto entre el valor real y los valores observados, otros astrónomos se centraron en minimizar la suma de los errores cuadráticos, de los cuadrados de los errores (los cuadrados se toman para dar el mismo valor a una discrepancia por defecto que por exceso), un método de estimación que en seguida se reveló como generalizable a más variables y más sencillo de cómputo que el que ideara Laplace. Era el método de mínimos cuadrados (figura 2). Este método fue dado a conocer por Adrien-Marie Legendre (1752-1833) en 1805, en su libro *Nuevos métodos para la determinación de las órbitas de los cometas*. Pero un joven matemático alemán, llamado Carl Friedrich Gauss (1777-1855), afirmó haber sido el primero en utilizarlo para predecir la órbita del asteroide Ceres, descubierto el primer día del siglo XIX, el 1 de enero de 1801.

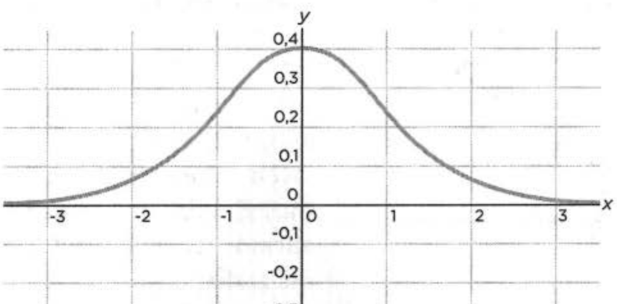
FIG. 2



En su obra *Teoría del movimiento de los cuerpos celestes* (1809), Gauss expuso, en el contexto de la teoría de errores, el método que había inventado en secreto para ajustar una curva dentro de una nube de puntos. Demostró que la distribución de los errores está relacionada con el método de mínimos cuadrados. Una vez determinada la curva que minimizaba el error cuadrático, Gauss observó que los errores cometidos en la aproximación se distribuían aleatoriamente alrededor de un valor medio. Esta distribución simétrica con forma de campana era la denominada *distribución normal* o *campana de Gauss* (figura 3), aunque en la época fue conocida simplemente como *ley del error*. Recíprocamente, Gauss demostró que si se suponía que los errores se distribuían de acuerdo con esta ley general, la función de mínimos cuadrados era la que minimizaba la probabilidad de error o, equivalentemente, la que hacía más verosímiles las observaciones (aunque en un primer momento no razonó así, sino que empleó el teorema de Bayes inspirándose en Laplace).

No mucho más tarde, Laplace importó los valiosos hallazgos del matemático alemán al dominio de la teoría de la probabilidad, añadiendo un resultado propio: el teorema central del límite, que afirma que si una medida es el resultado de la suma de un gran número de factores sometidos a error, esta se distribuirá normalmente con independencia de cómo lo haga cada uno de los factores en particular. Este teorema mostraba que la aproximación

FIG. 3



Ley de los errores según Gauss:

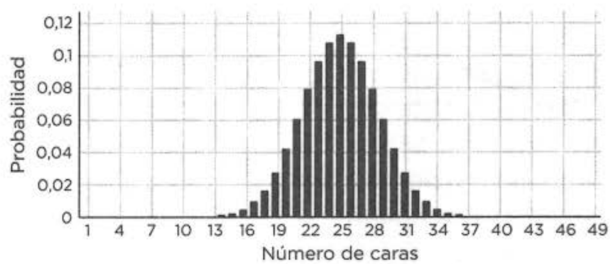
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

de la binomial a la normal, desarrollada por De Moivre como una herramienta de cálculo sin significado probabilístico, no era sino un caso particular de un resultado mucho más general. Cualquier suma o media, y no únicamente el número de éxitos en n experimentos (lo que había probado De Moivre), se distribuye aproximadamente como una normal si n es lo suficientemente grande (figura 4). En otras palabras, este teorema justificaba que, bajo ciertas condiciones muy generales, era plausible modelar una variable bajo estudio como si proviniese de una distribución normal. A este cúmulo de métodos y teoremas es a lo que los historiadores de la ciencia se refieren con la síntesis de Gauss-Laplace.

Si uno de los cursos progenitores de la estadística se encuentra en la francesa *Théorie mathématique des probabilités*, el otro hay que buscarlo en la «ciencia del Estado», es decir, en el análisis de datos socioeconómicos relacionados con el auge del comercio y los estados-nación. Con más precisión, en la confluencia de dos tradiciones iniciadas también a mediados del siglo XVII: la *Political Arithmetic* inglesa y la *Statistik* alemana. El término «aritmética política» fue introducido por William Petty, que pretendía operar sobre el cuerpo político imitando a la nueva filosofía natural, con el propósito de mejorar la toma de decisiones. Dentro de esta rama se encuentran las observaciones sobre tablas de mortalidad debidas a John Graunt en 1662, cuya indagación de estos datos demográficos era relevante para las rentas vitalicias y las primas de seguros. Es de destacar que estudiando estas tablas los hermanos Huygens entrevieron los juegos de azar como un modelo

La probabilidad de obtener un cierto número de caras al lanzar una moneda 50 veces presenta una distribución de probabilidad que se aproxima a la curva normal.

FIG. 4



EL PODER DE UN GRÁFICO ESTADÍSTICO

John Snow (1813-1858) fue un destacado médico inglés pionero en el dibujo de una suerte de pictograma orientado a demostrar que la virulenta epidemia de cólera que azotó Londres en 1854 se debía a un pozo de agua contaminada, alrededor del cual se acumulaban las víctimas (representadas por puntitos), y no, como era creencia habitual, por el contagio entre enfermos y sanos a través del aire. Las más de 700 personas que murieron en menos de una semana en el barrio del Soho lo hicieron porque todas ellas bebían de una fuente (marcada con un aspa en la calle Broad, en el centro de la imagen), contaminada con heces fecales. La ilustración corresponde al mapa original de John Snow. Los puntos representan las personas afectadas por la enfermedad, mientras que las cruces corresponden a los pozos de agua de los que bebían.



para inferir conocimiento acerca de otras porciones del mundo, y acuñaron el concepto de esperanza de vida a partir de la noción de esperanza o ganancia más probable de un juego. Por su parte, el término alemán *statistik* apareció en el contexto del interés por caracterizar a los nuevos estados —Prusia, en concreto— a través de sus estadísticas, de sus números e índices, puesto que los impuestos aduaneros entre los Estados alemanes se fijaban de conformidad con el número de habitantes de cada uno de ellos.

La tradición inglesa y la alemana convergieron hacia finales del siglo XVIII en las islas Británicas, pero no asimilaron las matemáticas francesas hasta bien entrado el siglo XIX. A partir de ese momento el estudio cuantitativo de la política y de la sociedad tomó prestadas las herramientas matemáticas de uso ya común en la doctrina del azar y la astronomía. La socialización de la teoría de probabilidades francesa se debe al astrónomo belga Adolphe Quetelet (como vemos, la conexión con la astronomía no es casual), aunque su lenta composición con la ciencia del Estado de raigambre inglesa y prusiana hubo de esperar a que tanto la obra de Laplace como la de Quetelet fuesen dadas a conocer en Gran Bretaña gracias al astrónomo John Herschel y al lógico Augustus de Morgan.

EL «HOMBRE MEDIO» DE QUETELET

Con la avalancha de números impresos que se produjo al final de la era napoleónica, el foco de las estadísticas pasó de ser el número de nacimientos, muertes y matrimonios al número de suicidios, asesinatos o analfabetos. Estas cifras relativas a la criminalidad y la educación fueron el caldo de cultivo en el que se engendró la idea del «hombre medio» *(homme moyen)*, que favoreció la erosión del determinismo.

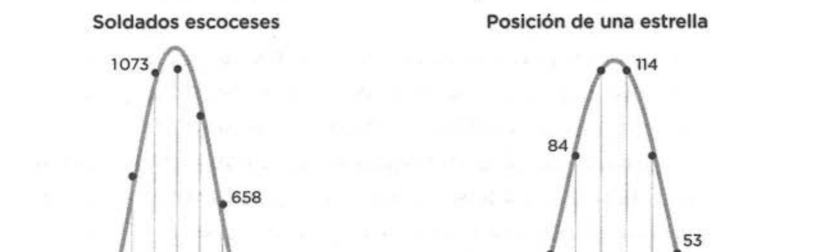
Adolphe Quetelet (1796-1874) completó sus estudios en París, donde a través de su maestro Joseph Fourier tomó contacto con la síntesis Gauss-Laplace. La perplejidad de Quetelet por las regularidades de la estadística surgió cuando, con el aumento de la burocracia, observó la terrible exactitud con que se producían los crímenes: las estadísticas criminales en Francia se sucedían

con valores anuales casi constantes. Entre 1825 y 1830 el número anual de acusados estaba siempre alrededor de 7 100, y el de condenados, en torno a 4 400. A su regreso a Bruselas se interesó por el planteamiento de censos y encuestas.

Inicialmente, llevado por su deseo juvenil de ser escultor, Quetelet aplicó las nociones probabilísticas que manejaba con soltura en astronomía y geodesia a la medición del cuerpo humano (al astrónomo belga le debemos la definición del índice de masa corporal que determina la obesidad). En 1835 anunció que la ley del error —o «ley de las causas accidentales», como prefería denominarla— se aplicaba a las características humanas, físicas y de comportamiento, siendo el concepto central el de *promedio*, pues el valor medio de la distribución de la característica bajo estudio representaba al «hombre medio». Ciertas mediciones antropométricas, como la estatura de los reclutas franceses o el tórax de los soldados escoceses, se distribuían aproximadamente como la curva acampanada de Gauss. En efecto, en 1845, tras tabular y representar los datos relativos a los perímetros de pecho de 5 738 soldados escoceses, tomados de una revista médica de la época, observó el parecido entre la curva de frecuencias resultante y la que aparecía a la hora de medir la posición de una estrella (figura 5).

A la izquierda, curva de frecuencias correspondiente a la amplitud de pecho de 5 738 soldados escoceses según Quetelet (1845). A la derecha, curva de frecuencias de los errores cometidos en la observación de una estrella según el astrónomo Friedrich W. Bessel (1818).

FIG. 5



Pero mientras que el astrónomo medía muchas veces la misma estrella, existiendo un valor real de la posición, Quetelet mostraba datos de distintos soldados y detrás de su curva no había un valor real del perímetro de pecho. Quetelet argumentó que medir el perímetro de pecho de muchos soldados era como medir muchas veces el perímetro de pecho de un mismo soldado, del «soldado medio». Y, dando un enorme salto ontológico, propuso que la razón es que la naturaleza apunta a una especie de hombre promedio, y que los que están en los extremos de la campana son desviaciones azarosas del canon ideal. Su obra marcó el inicio de la física social y sirvió de propaganda internacional del valor de las estadísticas, catalizando la formación de la Sociedad Estadística de Londres, entre otras instituciones estadígrafas.

No obstante, no hay que olvidar que la conexión de la probabilidad y la estadística con la sociedad ya estaba de forma embrionaria en Laplace, puesto que el astrónomo francés recogió el testigo de la «aritmética moral» esbozada por Condorcet en su *Ensayo sobre la aplicación del cálculo a la probabilidad de las decisiones* (1785), cuya meta puede retrotraerse, a su vez, a la última parte del tratado de Bernoulli, que estaba dedicado a la aplicación del cálculo de probabilidades a cuestiones civiles, morales y económicas, buscando aunar la sabiduría del filósofo con la prudencia del político, según sus propias palabras. En el popular *Ensayo filosófico sobre las probabilidades*, publicado originalmente como introducción a la segunda edición de la *Teoría analítica de las probabilidades* (1814), Laplace dejó escrito que «los problemas fundamentales de la vida no son en el fondo más que problemas de probabilidades». No era un simple matrimonio de conveniencia. Para Laplace la probabilidad era la base de la inferencia científica, de la teoría del error, de la filosofía de la causalidad y, atención, de la cuantificación de la credibilidad de los testimonios. Si el cálculo de probabilidades se había revelado tan eficaz en las ciencias naturales, ¿por qué no iba a serlo también en las ciencias políticas y morales? En su opúsculo, Laplace equiparaba las decisiones de una asamblea o las sentencias de un tribunal con las posibles bolas que podían extraerse de una urna, a fin de determinar la probabilidad de error en función del número

de diputados que formaran la asamblea o del número de votos que hiciesen falta para condenar al acusado, perfeccionando así los cálculos al respecto que hiciera Condorcet antes de la Revolución. No deja de tener su gracia, como no dejó de advertir Laplace, que una ciencia que comenzó con consideraciones sobre monedas, dados y barajas se convirtiera pasado el tiempo en uno de los objetos más importantes del conocimiento humano.

> «La urna a la que interrogamos es la naturaleza.»
>
> — Adolphe Quetelet (1845).

De hecho, Siméon Denis Poisson, el discípulo más prometedor de Laplace, contribuyó significativamente a la orientación social que tomó la estadística con Quetelet. En 1835, mientras trabajaba en cuestiones de matemática electoral y jurisprudencia, formuló la «ley de los grandes números», que proveyó una mejor base para aplicar la matemática de las probabilidades a los problemas sociales, explicando la estabilidad estadística a través de los cambios sociales. Grandes números de individuos, actuando independientemente en un sistema, producen regularidades que no dependen de su coordinación mutua, de manera que es posible razonar sobre la colectividad sin ningún conocimiento detallado de los individuos. En consecuencia, no se podía predecir el comportamiento particular de un individuo, pero sí el comportamiento promedio de la población. Se trataba de otra manifestación más de la regularidad estadística del mundo. Poisson y Quetelet eran dos astrónomos que veían en la conducta y en las características de sus millones de conciudadanos regularidades dignas de los astros.

En suma, Quetelet partió de la curva de Gauss, deducida previamente como ley del error o como distribución límite en juegos de azar como el lanzamiento de monedas, y aplicó esta misma curva a fenómenos biológicos y sociales donde la media no es una magnitud real, transformándola en una cantidad real. La media no era un rasgo de un individuo concreto, sino una característica de la población que simplificaba los datos de partida. Servía para representar a la población en el carácter bajo estudio, de manera que los diversos individuos se mostraban como desviaciones mayores o

menores de este valor, del hombre medio. Para Quetelet, las variaciones observadas eran simples perturbaciones, errores naturales. Desinteresándose por el estudio intrínseco de la variabilidad, el astrónomo belga identificaba la media con lo justo y lo correcto. Con la recepción de sus trabajos en Inglaterra la curva acampanada fue rebautizada como *ley normal*. Las personas normales eran aquellas que se ajustaban a la tendencia central de las leyes sociales que cuantificaban la estatura, el peso o la inteligencia. La sociología proseguiría en esta dirección al catalogar a aquellas personas cuyos valores se encontraban en los extremos como patológicas, «anormales». Pero la influencia de la obra de Quetelet no se detiene aquí, pues puso a James Clerk Maxwell (1831-1879) en el camino de la mecánica estadística: las moléculas de un gas son como los individuos de una población, ya que el desorden a escala individual se transforma en un orden a escala poblacional. No en vano, la teoría de gases fue la otra materia —junto con la teoría de errores— que permitió a Fisher aprender los métodos estadísticos clásicos.

SIR FRANCIS GALTON, EL «HOMBRE MEDIOCRE» Y LA EUGENESIA

Para comprender cómo los métodos estadísticos pasaron del campo de la física social al campo de la antropología física y, en especial, a la biología evolutiva, hay que atender al cambio en el estudio de la variabilidad estadística que propició la aparición del darwinismo y la eugenesia. Fue la insuficiencia de las teorías genéticas de Charles Darwin (1809-1882) lo que animó a Francis Galton (1822-1911), de facto su primo, a tratar de resolver los problemas de la herencia mediante el análisis matemático que los datos biológicos demandaban.

Galton, que nació el mismo año que Gregor Mendel (1822-1884), era trece años más joven que Darwin. Tras estudiar medicina y matemáticas gracias a la generosa herencia paterna, se embarcó hacia África como explorador (entre otros inventos, como los mapas anticiclónicos, patentó el saco de dormir). A su vuelta



Este retrato, de 1913, muestra a un joven Fisher graduado en Matemáticas tras su paso por la Universidad de Cambridge, donde creció su interés por la genética y la evolución a raíz de la lectura de una serie de artículos de Karl Pearson.

a Inglaterra, coincidiendo con la consolidación de la antropología colonialista, se interesó por la evolución. Galton quedó cautivado por la lectura del primer capítulo de *El origen de las especies* (1859), que aborda la variación bajo domesticación, relativa a la cría de animales, y en seguida estableció una correspondencia regular con Darwin que duraría hasta la muerte de este último. Barajando la posibilidad de dirigir de forma controlada la selección natural de la especie humana, Galton comenzó a pensar seriamente en la mejora de la humanidad a través de la crianza selectiva de los seres humanos. En *Genio hereditario* (1869), decía:

> De la misma manera que se logra una raza permanente de perros o caballos dotada de especiales facultades para correr o hacer cualquier otra cosa, sería factible producir una raza de hombres altamente dotada mediante matrimonios sensatos durante varias generaciones consecutivas.

En 1883, Galton acuñó, precisamente, el término *eugenesia* («ciencia de la mejora de la raza»). Este concepto haría fortuna en la sociedad británica finisecular, preocupada por el declinar de su imperio tanto en el exterior (frente a otros imperios) como en el interior (con el avance de las clases bajas, del lumpemproletariado, cuyo índice de natalidad era muy superior al de la clase alta). Y arraigaría en Estados Unidos y en la Alemania nazi, con la promulgación de leyes de esterilización forzosa para enfermos mentales e indigentes. El movimiento eugenésico prácticamente no se aplacaría hasta que se apagasen los hornos crematorios en Centroeuropa y se proclamara la división humana en razas como un mito propio de la antropología física prebélica.

Galton creía firmemente que la población inglesa estaba sufriendo una suerte de involución, una degeneración biológica que se transmitía hereditariamente y que se manifestaba en las dificultades militares que atravesaba el Imperio británico, achacables en su opinión a la creciente debilidad innata de las tropas. La ciencia eugenésica debía aportar la solución al problema favoreciendo que las mejores estirpes se reprodujesen y limitando la procreación de las capas más desfavorecidas.

A diferencia de Galton, Darwin mantenía una actitud más prudente. En *El origen del hombre y la selección en relación al sexo* (1871), abordó la cuestión de las razas humanas y, aunque aceptó las teorías eugenésicas, expresó ciertas reservas. Puede parecer sorprendente que Darwin aceptara estas teorías basadas aparentemente en la herencia de los caracteres adquiridos popularizada por Lamarck, pero la explicación del mecanismo hereditario detrás de las adaptaciones era una anomalía recurrente para el darwinismo clásico. La teoría de la «pangénesis», propuesta por Darwin a falta de otra mejor, era totalmente compatible con la herencia lamarckiana (aunque Galton difundió, para enfado de Darwin, los resultados de una serie de experimentos con conejos que contradecían la existencia de «semillas sanguíneas»). Solo el neodarwinismo, resultante de la síntesis del darwinismo clásico con la genética mendeliana y poblacional, expulsó al lamarckismo de la escena científica (la causa de las variaciones hereditarias son las mutaciones en el ADN).

Hacia el final de su vida, Galton incluso escribió una novela utópica, titulada *Kantsaywhere*, sobre una sociedad que vivía feliz bajo preceptos eugenésicos dictados por sacerdotes-científicos, que su sobrina (Galton no tuvo hijos en su matrimonio), irritada por algunas escenas subidas de tono, quemó parcialmente. La influencia de las ideas galtonianas fue notable, dando alas al darwinismo social y a la introducción de la estadística en el estudio de la psicología. Los test antropométricos de Galton se transformaron a la vuelta de siglo en los célebres test de inteligencia.

LA LEY DE REGRESIÓN A LA MEDIA Y LA NOCIÓN DE CORRELACIÓN

La contribución más duradera de Galton fue la utilización de la estadística como herramienta destinada a domesticar la variabilidad biológica hereditaria. Para el polivalente científico inglés era un dogma que uno solo conoce una cosa cuando puede medirla, lo que a la postre significó la consagración de la antropología física

cuantitativa o antropometría. A juicio de Galton, las características físicas, tales como la altura, el peso o los rasgos de personalidad, son heredadas. Galton creía que la unión de dos personas inteligentes produciría una persona más inteligente, del mismo modo que la unión de dos personas altas produciría otra persona más alta. Sin embargo, los experimentos sobre la herencia que realizó a lo largo de su vida le llevaron a descubrir una nueva regularidad estadística, distinta de la esperada, y que denominó *reversión a la mediocridad* —más tarde *regresión a la media*— en su libro *Herencia natural* (1889). Galton empleó este concepto para designar la relación que existía entre la estatura de padres e hijos. Observó que si los padres son altos, los hijos generalmente también lo son, y si los padres son bajos, los hijos son también de menor estatura. Pero cuando el padre es muy alto o muy bajo, aparece una apreciable regresión hacia la estatura media de la población, de modo que los hijos retroceden o regresan hacia la altura media de los padres. Galton extendió este resultado planteando una ley universal sobre la herencia ancestral: cada peculiaridad en un hombre es compartida por sus descendientes, pero en media en un grado menor (hoy se sabe que más que una regularidad biológica se trata de una regularidad puramente estadística, debida al azar: lo más probable es que las realizaciones de una variable aleatoria normal sean próximas a su media o valor esperado).

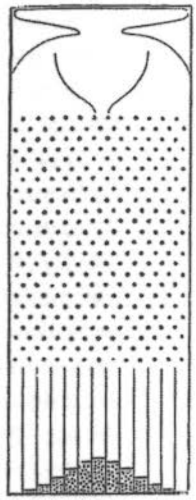
Hacia 1877 Galton había descrito este mismo fenómeno experimentando con el tamaño de las semillas de generaciones sucesivas de guisantes. Mientras Mendel experimentaba con caracteres cualitativos (color, rugosidad, etc.) de los guisantes, Galton lo hacía con caracteres cuantitativos (tamaño, diámetro). Cuando repitió su estudio con registros antropométricos (donde, por cierto, introdujo el uso de los percentiles y revalorizó el uso de la mediana y los cuartiles), observó con algo de ayuda la siguiente relación lineal:

$$\text{Altura del hijo (en cm)} = 85 \text{ cm} + 0{,}5 \cdot \text{Altura del padre (en cm)}.$$

Se trataba de una de las rectas de regresión. Además, conjeturó que la intensidad de la relación entre las dos variables —la al-

EL *QUINCUNX*

El polifacético Galton buscaba explicar el hecho de que ciertas medidas físicas (como la altura de las personas o el diámetro de los guisantes) se distribuyen normalmente. Para argumentar que la ley normal era la ley de la genuina variación y no solo la ley del error, ideó en 1873 el *quincunx*, un dispositivo cuyo nombre proviene de los sembrados en que cada árbol está rodeado por otros cuatro árboles, y que sirve para ilustrar el teorema central del límite. El dispositivo consiste en un tablero en el que se introducen unos guisantes a modo de bolitas por el extremo superior, que van cayendo rebotando de manera azarosa en los «árboles» hasta ser recogidos en unos compartimentos separados en el otro extremo. Con este dispositivo, Galton demostró que las bolitas dibujan en el extremo inferior una campana de la distribución normal, como se observa en la ilustración. Mediante este ingenioso mecanismo explicaba la prevalencia de la distribución normal e, incluso, ilustraba la herencia mediante una disposición en fases. Interrumpiendo el paso de las bolitas en alguna zona, para representar las influencias dominantes en la herencia, observó que aún se dibujaba una curva normal, aunque más pequeña y menos dispersa. El científico inglés era verdaderamente un genio en cuanto a transformar representaciones abstractas en modelos físicos. Con su investigación, reconcilió la teoría de errores —según la cual una acumulación de desviaciones accidentales da lugar a una distribución normal— con la herencia, que si bien tiene desviaciones accidentales, también contiene obvias correlaciones, ya que cada organismo tiende a semejarse a sus ancestros.



Esquema del *quincunx* en el libro *Herencia natural* (1889).

tura del padre y la del hijo— podía cuantificarse numéricamente. Era la mayor innovación estadística de la centuria: la correlación.

Mientras que la obra de Galton sobre la regresión fue el resultado directo de sus investigaciones sobre la herencia, su teoría de la correlación nació de los problemas de identificación de criminales (un tema en el que fue pionero al introducir el uso de las huellas dactilares). Galton comprendió en seguida que en el sistema de identificación propuesto por el policía francés

Alphonse Bertillon (1853-1914) había mucha redundancia. Bertillon registraba la estatura, las dimensiones de los pies, de los brazos y de los dedos de cada persona; pero estas cuatro medidas no eran independientes entre sí, pues las personas altas suelen tener los pies, los brazos y los dedos largos. Galton conjeturó que en esencia se trataba de la misma cuestión que había rozado en su estudio de la regresión: la correlación entre variables. En un artículo firmado en 1888, introdujo una primera medida matemática de la correlación, es decir, del grado de dependencia entre variables, aunque la definición como coeficiente vendría de la mano del economista Francis Y. Edgeworth en 1892 y sería redondeada por el matemático Karl Pearson —a quien presentaremos en el próximo capítulo— en 1896, que otorgaría parte del prestigio por el descubrimiento al astrónomo francés Auguste Bravais, que ya en 1846 había dado una formulación matemática similar a la hora de estudiar los errores correlativos entre las coordenadas de posición de un objeto. Hoy en día se lo conoce como *coeficiente de correlación lineal de Pearson*, y permite estudiar correlaciones positivas y negativas (un caso que Galton no pareció plantearse, cuando el incremento en la primera variable se traduce en un decremento en la segunda).

Galton siempre rememoraba que la eugenesia, el deseo de mejorar las cualidades raciales físicas o mentales, fue el impulso que le empujó a estudiar el problema colateral de la variación estadística. Hasta entonces, los métodos estadísticos solo se preocupaban por los promedios colectivos, desinteresándose por las variaciones individuales. Para Quetelet, el hombre medio era el centro de gravedad del cuerpo social, alrededor del cual oscilaban los átomos sociales, los hombres particulares. Este hombre medio era el canon de perfección, pues estaba libre de excesos y defectos. Galton reconocía su deuda con Quetelet al referirse a él como la mayor autoridad en la estadística social, por cuanto difundió el uso de la curva normal, no como ley del error, sino como descripción de la distribución de las mediciones. Pero entre ambos científicos se produjo una transición fundamental en la concepción de las leyes estadísticas, debida en gran parte a la fascinación de Galton con lo excepcional, en oposición a la preocupación de Quetelet por los promedios.

Mientras que Quetelet pensaba en la tendencia central y, por tanto, en la media, Galton, siempre preocupado por la excepción, se fijaba en las colas de la distribución y en la dispersión. Galton atendía a aquellos individuos que se desviaban ampliamente de la media por exceso o por defecto: el hombre medio de Quetelet ya no era el prototipo de perfección, sino un hombre mediocre que necesitaba evolucionar. Lo excelso se encontraba en uno de los extremos de la curva normal del talento. Este cambio revolucionario solo fue posible cuando la normalidad devino mediocridad gracias a que la selección natural de Darwin y, de forma asociada, la reforma eugenésica resucitaron el interés por la variabilidad: las características excepcionales ya no eran errores de la naturaleza, desviaciones del hombre medio ideal, sino variaciones importantes para la mejora de la raza. La estadística pasó de ser una herramienta concebida para reducir el error a un modelo para representar la variación debida al azar. La reinterpretación de la curva normal como la ley de la genuina variación, en vez que del mero error, fue el resultado central del pensamiento estadístico del siglo XIX.

«La ley normal habría sido deificada por los griegos, si la hubieran conocido.»

— Francis Galton, *Herencia natural* (1889).

En resumen, nuestro protagonista, Fisher, conoció los entresijos de la estadística gracias a un curioso maridaje de saber astronómico, físico y natural. A través de la teoría astronómica de los errores, asimiló la síntesis Gauss-Laplace, en otras palabras, la yuxtaposición entre el cálculo de probabilidades, el método de mínimos cuadrados y la ley del error. Por medio de la teoría cinética de los gases, aprendió a modelar colectividades mediante la distribución normal. Y, finalmente, los avances en biología y antropología auspiciados por Galton le permitieron cobrar contacto con la principal novedad estadística decimonónica: la correlación.

CAPÍTULO 2

Karl Pearson y la escuela biométrica

La obra de Fisher no puede entenderse sin contrastarla con la de su inmediato predecesor, Karl Pearson. En su intento por desarrollar una teoría matemática de la evolución, Pearson alumbró algunos de los conceptos y métodos estadísticos clásicos. Entre los primeros, están los histogramas y la desviación típica. Entre los segundos, el análisis de la regresión y el test de la χ^2. Las rectificaciones que el joven Fisher haría a varios trabajos de Pearson conducirían a una enconada rivalidad de por vida entre los dos.

Durante su estancia en Cambridge, Fisher leyó los artículos publicados por el matemático Karl Pearson bajo el sugestivo título de *Contribuciones matemáticas a la teoría de la evolución*. Instigado por la lectura de esta serie de artículos que conjugaban sus dos aficiones principales (la estadística y la biología), Fisher realizó su primera investigación científica original. Lo hizo en 1912, con solo veintidós años de edad y sin haber terminado aún los estudios.

Al dejar la universidad, las finanzas familiares no estaban demasiado boyantes y Fisher no tardó en buscar una ocupación como estadístico en una compañía mercantil e, incluso, trabajar durante un tiempo en una granja en Canadá. En 1914, de regreso a Inglaterra, coincidiendo con el estallido de la Primera Guerra Mundial, trató de alistarse, pero le declararon no apto para el servicio militar por culpa de su vista maltrecha. En 1917 contrajo matrimonio en secreto con Ruth Eileen (que, entonces, contaba con diecisiete años), con la que tendría ocho hijos, dos niños y seis niñas (una de ellas, Joan, la mayor, se casaría con el también estadístico George E.P. Box). En 1919, tras ejercer como profesor de Física y Matemáticas en varias escuelas, llegó su gran oportunidad, y lo hizo por partida doble. Pearson le ofreció una plaza como estadístico en el Laboratorio Galton y, simultáneamente, le ofrecieron otra en la Estación Agrícola Experimental

de Rothamsted, el instituto de investigación agrónoma con más tradición del Reino Unido.

Fisher resolvió el dilema inclinándose por la segunda opción, por Rothamsted. La razón principal fue que trabajar en el Laboratorio Galton conllevaba que Pearson tenía que supervisar sus publicaciones, una condición que no estaba dispuesto a aceptar. Ni mucho menos. Sobre todo cuando los puntos de fricción entre ambos se habían ido acumulando durante los últimos años y seguirían haciéndolo: la distribución *correcta* del coeficiente de correlación, el número *exacto* de grados de libertad en el test de la χ^2 («chi-cuadrado»), la *eficiencia* del método de estimación de los momentos... Lo que había comenzado siendo una relación amistosa, acabó enturbiándose a causa de varios malentendidos. Pese a su juventud, Fisher corrigió el trabajo de Pearson y de sus colaboradores más cercanos en varios aspectos, un hecho que el segundo no terminó de encajar nunca, aunque desde luego el carácter altivo que destilaba Fisher no ayudó a mejorar las cosas. Para poder explicar en qué sentido los errores teóricos de Karl Pearson impulsaron el despegue de la investigación de Fisher, además de precipitar la abrupta ruptura entre ambos, es preciso acercarnos a la figura principal de la estadística victoriana y su magna obra.

ENTRE LA ELASTICIDAD Y LA BIOMETRÍA

A partir de 1884 Pearson fue profesor de Matemática aplicada y Mecánica en el University College de Londres. Tras acceder a la cátedra, se había especializado en teoría de la elasticidad, ya que en la segunda mitad del siglo XIX la elasticidad era el problema por excelencia de la cosmología, puesto que la trasmisión electromagnética precisaba de un éter elástico. Pero Pearson poseía una vocación no estrictamente científica. Gran parte de su magnetismo personal provenía de su enérgico diletantismo humanista, un gusto por la literatura, la historia o la filosofía que ni siquiera cesó cuando se concentró en el cultivo de técnicas

UN PERSONAJE IMPROBABLE



Karl Pearson (1857-1936) vino al mundo en el seno de una familia londinense que pertenecía a la clase media profesional, lo que le permitió graduarse en Matemáticas en Cambridge en 1879 y realizar estudios de posgrado en las universidades de Heidelberg y Berlín, donde leyó y escribió incansablemente sobre múltiples temas: poesía, teatro, ética, socialismo, derechos de la mujer, etcétera, y hasta llegó a escribir un drama, *El nuevo Werther*, publicado bajo el pseudónimo de *Loki* en 1880. En 1892, Pearson publicó *La gramática de la ciencia*, un libro que recogía su filosofía de la ciencia, en la que se mezclan el idealismo aprendido del filósofo neokantiano Kuno Fisher y el positivismo expuesto por Ernst Mach, que hizo suyos en Alemania (no en vano, Pearson cambió la C de su nombre de pila por una K tras su estancia). Este libro conoció varias ediciones en vida del autor, gozando de gran éxito. Albert Einstein, por ejemplo, formó un pequeño grupo de lectura del mismo en Berna hacia 1902, y su contenido llegó a influirle en la formulación de la teoría de la relatividad especial. Una de las ideas centrales del libro es que la función de la ciencia debe limitarse a describir los hechos observables, evitando cualquier clase de recaída en la metafísica. Las leyes científicas no son explicaciones causales, sino resúmenes ordenados de los fenómenos. En otras palabras, no nos explican por qué suceden las cosas, sino que simplemente describen cómo lo hacen. Pearson quería promover científicamente el bienestar nacional y mantenía que la Ciencia, con mayúscula, tenía que convertirse en la base cultural común de la civilización. Además, al igual que Galton, defendía las bondades de la eugenesia, manifestando en varias ocasiones su deseo de que aquellos miembros de la comunidad que presentasen una gran desviación física o mental respecto de la media tuviesen una selección sexual más cuidadosa. El científico inglés se mostraba preocupado por el declinar de la nación británica como consecuencia, en su opinión, de la disminución de la fertilidad en las clases liberales. Pero su creencia en la eugenesia científica se combinaba con una defensa ardorosa del socialismo. En la lucha darwinista por la existencia entre las naciones, el socialismo parecía imponerse como una lección histórica.

estadísticas dentro del dominio de la biología evolutiva. Sin ir más lejos, en *El nuevo Werther*, obra que Pearson publicó en 1880, exclamaba:

> Los gigantes de la literatura, los misterios del espacio multidimensional, los intentos de Boltzmann y Crookes por escudriñar el laboratorio de la naturaleza, la teoría kantiana del universo y los últimos descubrimientos en embriología, con las maravillosas aventuras sobre el desarrollo de la vida... ¡qué inmensidad más allá de nuestro entendimiento!

La metamorfosis de este matemático experto en teoría de la elasticidad en el primer estadístico en sentido moderno no se puede explicar si no se tiene en cuenta que se trataba de un prodigioso pero anacrónico científico renacentista, obsesionado con la persecución de la verdad numérica y espiritual. No es casual que una de las metas a las que aspiraba Pearson fuese que los futuros estadísticos aunasen las dos culturas (las ciencias y las letras), interesándose tanto por la resolución de problemas como por la historia de la disciplina, a la manera que él mismo escribió, en sus tiempos mozos, una historia cronológica de la teoría de la elasticidad y, ya en su madurez, una ambiciosa biografía en tres volúmenes de su admirado Francis Galton, así como una colección de lecciones sobre los orígenes de la estadística en relación con el pensamiento religioso.

Hacia 1892 se produjo un cambio drástico en los intereses científicos de Pearson. Por medio de la amistad con Walter Frank Raphael Weldon (1860-1906), profesor de Zoología en el University College, a quien había conocido un año antes en una reunión para reformar la universidad, se interesó por el desarrollo de métodos estadísticos que permitieran avanzar en el estudio de la herencia y la evolución, ya que después de la muerte de Darwin se trataba —con la notable excepción de las investigaciones de Galton— de un campo prácticamente moribundo. Es de destacar que Pearson había regresado de su viaje formativo por tierras alemanas convertido no solo en un ferviente socialista, sino en especial en un darwinista convencido, ya que había asistido a las

clases de Emil du Bois-Reymond, hermano del matemático Paul du Bois-Reymond, en Berlín.

Raphael Weldon precisaba de ayuda con el análisis de los datos zoométricos recolectados con el propósito de esclarecer cómo operaba la selección natural, que constituía su hipótesis de trabajo. En 1890 había demostrado, basándose en mediciones realizadas en *Decapod crustacea* (una especie de cangrejo), que la distribución de las variaciones en este animal era casi la misma que la observada por Quetelet y Galton en el hombre: la ley normal. Era la primera vez que las técnicas estadísticas desarrolladas por Galton en el ámbito de la antropología se aplicaban a la biología. Por vez primera se calculaba también un coeficiente de correlación orgánico, entre los tamaños de dos órganos. Galton, que leyó la memoria en calidad de árbitro, no tardó en establecer contacto con Weldon, que en sus estudios con cangrejos se había convencido de que la evolución era en el fondo un problema estadístico. Los dos mecanismos de la teoría de la evolución, la producción de variabilidad y la selección natural mediante la lucha por la existencia, tenían un innegable atractivo desde este punto de vista. La producción de variabilidad entroncaba con el azar, con el cálculo de probabilidades; la selección natural, con el estudio de poblaciones, ya que son las unidades que van a sufrir la evolución en su conjunto. Por este motivo, Weldon necesitaba la colaboración urgente de un colega matemático.

Con treinta y cinco años cumplidos, Pearson comenzó a estudiar los métodos estadísticos tal y como estos aparecían en muchos manuales continentales dedicados a la demografía. Asimismo, releyó los libros de Galton (a quien conoció en persona en 1894 por mediación de Weldon), ya que su primera lectura de *Herencia natural* (1889) no había sido muy positiva, a tenor de la opinión que expresó en el londinense Club de Hombres y Mujeres del que era miembro:

> Personalmente debo decir que existe un considerable peligro en aplicar los métodos de las ciencias exactas a los problemas de la ciencia descriptiva, tanto si se trata de problemas de la herencia como de política económica.

Es más, en el ejemplar conservado del libro de Galton, Pearson dejó constancia autógrafa de su exasperación por los argumentos expresados por su autor: a su juicio se trataba de meras analogías sin valor científico alguno. Pese a todas estas evidencias, sigue leyéndose demasiado a menudo que el ímpetu estadístico de Pearson radicó en la lectura del libro de Galton, de quien se le considera erróneamente discípulo. Probablemente, Weldon fue el responsable de su cambio de opinión, dado que consiguió ilustrar con ejemplos concretos cómo las técnicas estadísticas planteadas por Galton podían aplicarse con acierto al material biológico.

Según reinterpretó años después su acercamiento a la obra de Galton, Pearson quedó sorprendido por un descubrimiento del eminente científico: había una categoría más amplia que la causalidad, a saber, la correlación, de la cual la causalidad era solo el límite. Gracias a esta nueva concepción, la sociología, la psicología, la antropología y la biología podían entroncar con las matemáticas. Mientras que el físico piensa que un cierto valor de x produce (causa) un valor determinado de y, el estadístico cree que la relación entre x e y es más vaga, meramente probabilitaria. Galton liberó a Pearson del prejuicio de que las matemáticas solo podían aplicarse a los fenómenos naturales bajo la categoría causal. No cabe duda de que su renovada fascinación con la obra de Galton se debió en parte a su interés compartido por la eugenesia.

La voluntad de investigar conjuntamente determinó la fundación de la Escuela Biométrica por Weldon y Pearson bajo la influencia directa de Galton en 1893. El término *biometría* fue acuñado, precisamente, por Pearson con el significado de «ciencia de la medida de la vida». La escuela puso las bases de la estadística matemática entre 1895 y 1915, aun cuando la mayoría de edad de la disciplina no llegó hasta el período que va de 1915 a 1935, capitaneado por Fisher. En ambos casos, fue la necesidad de resolver problemas biológicos —relacionados, durante el primer período, con la herencia y la evolución, y, en el segundo, con la genética y la experimentación agrícola— lo que aceleró la cristalización de nuevas herramientas estadísticas.

CONTRIBUCIONES MATEMÁTICAS A LA TEORÍA DE LA EVOLUCIÓN

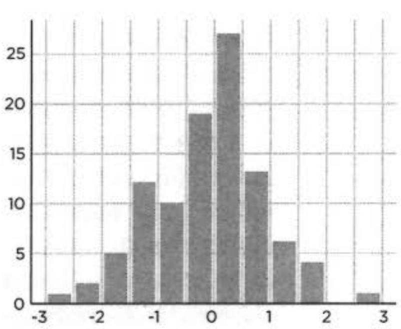
Con extraordinaria celeridad, Pearson empezó a producir nuevos conceptos y métodos, que muy pronto se revelaron como indispensables para cualquier aplicación de la estadística en otro campo. Antes de darlos a la imprenta, Pearson presentó muchas de sus ambiciosas ideas en una serie de conferencias vespertinas que impartió entre 1891 y 1894 en el Gresham College. Las primeras ocho conferencias cubrieron aspectos básicos de la filosofía de la ciencia, que fueron recogidos en el libro *La gramática de la ciencia* (1892). En la edición de 1900 de esta obra, escribía:

> De la misma manera que podemos predecir poco o nada de un átomo individual, poco podemos predecir de una unidad vital individual. Solo podemos manejar las estadísticas de su conducta promedio. Pero tenemos unas leyes de la variación y de la herencia casi tan definitivas y generales como las leyes de la física.

Las treinta conferencias restantes se dedicaron por completo a la «geometría de la estadística» y la «geometría del azar», por emplear los rótulos originales. El matemático inglés eligió estos tópicos porque muchos de los asistentes como público trabajaban por las mañanas en el distrito financiero de la City y pensó, no sin razón, que presentar la estadística mediante gráficos e ilustraciones podía ser de su agrado. En una de estas conferencias introdujo, por ejemplo, los *histogramas* (figura 1), un diagrama que podía ser de utilidad en historia —como su nombre quería indicar— para representar la evolución del número de habitantes o de los ingresos de un reino mediante intervalos de tiempo que estarían adyacentes unos con otros. Estas lecciones marcaron el comienzo de una nueva época en la

En los histogramas, a diferencia de en los diagramas de barras (que se usan para reflejar datos no agrupados), las clases no aparecen separadas sino contiguas.

FIG. 1



teoría y en la práctica de la estadística. No por casualidad, Pearson afirmó ante los presentes que a esta ciencia le aguardaba un futuro prometedor, pues daría lustre a otras ramas de la matemática e incluso al estudio de la biología.

Uno de las primeros conceptos que forjó fue el de «desviación típica» (o «desviación estándar»), que a partir de 1893 sustituyó al de «error probable», introducido por el astrónomo Friedrich W. Bessel alrededor de 1815, como más adecuado para medir la variación biológica. Mientras que la mayoría de los matemáticos y astrónomos del siglo XIX se habían orientado al estudio de medidas de la concentración y de la posición de los datos, Pearson se preocupó por medir su dispersión o variabilidad. Si Quetelet revalorizó el uso de la media y Galton hizo lo propio con la mediana —una medida propuesta por Antoine Augustin Cournot—, los cuartiles y los percentiles, Pearson bautizó a la raíz cuadrada del promedio de los cuadrados de las diferencias de cada dato respecto de la media (una expresión conocida en la época como «error cuadrático medio») con el nombre de *desviación típica* y el signo σ, para subrayar que la variación no tenía por qué interpretarse siempre como un error.

El error probable quedaba caracterizado porque dividía las posibles observaciones de un astro —distribuidas según la curva gaussiana en torno al valor real— en dos clases igualmente probables: a largo plazo, la mitad de las observaciones caerían en un entorno de su media aritmética de radio el error probable, y la otra mitad caería fuera, fallando demasiado por exceso o por defecto. El error probable representaba lo que hoy a veces se denomina *desviación absoluta respecto de la mediana*. La desviación típica de una serie de observaciones se calculaba más fácilmente y poseía mejores propiedades: la desviación típica de una distribución de error teórica, de un modelo de probabilidad, no era más que la versión continua de la fórmula discreta antes enunciada. En la distribución normal el error probable es de 0,6745 veces la desviación típica, de manera que mientras que en un entorno de la media de radio del error probable cae el 50% de las observaciones, en un entorno de radio de la desviación típica cae aproximadamente el 68%, y en un entorno de dos desviaciones típicas, algo más del

FIG. 2

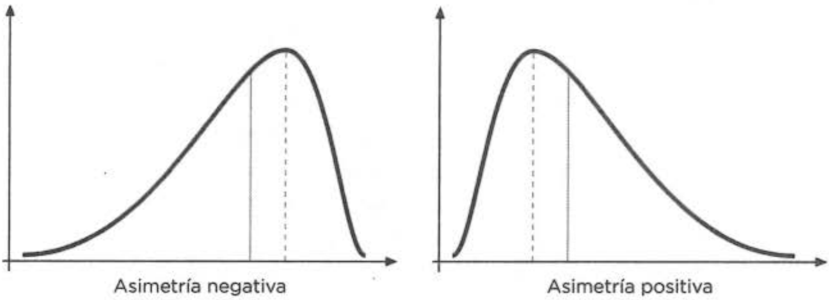


FIG. 3

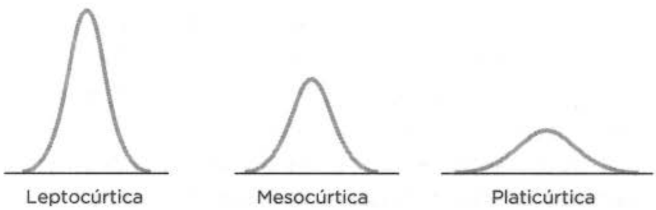


FIGURA 2:
A la izquierda, una distribución con asimetría negativa; a la derecha, con asimetría positiva. En trazo continuo, la media; en trazo punteado, la moda (el signo de la diferencia entre estos dos valores permitía a Pearson establecer el tipo de asimetría).

FIGURA 3:
La curtosis (término derivado de la palabra griega que significa curvado o arqueado) mide el grado de apuntamiento de una distribución en comparación con la distribución normal, definida como mesocúrtica.

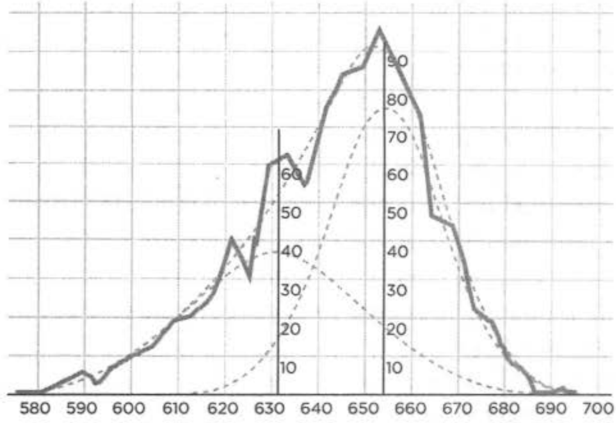
95% (si la distribución no es normal solo puede asegurarse que entra al menos el 75% de las observaciones).

Además, Pearson ideó el coeficiente de variación, definido como el cociente de la desviación típica y la media en valor absoluto, que servía para comparar la variabilidad entre distintos conjuntos de datos, midiendo en cada uno de ellos el grado de representatividad de la media, esto es, si los datos están o no concentrados alrededor suyo. Finalmente, ideó otras dos medidas descriptivas, el coeficiente de asimetría (figura 2) y el coeficiente de apuntamiento o *curtosis* (figura 3) para medir la forma de una distribución: si es simétrica o asimétrica respecto de la media, y si es más apuntada o más achatada que la distribución normal. En suma, Pearson inventó toda una colección de medidas realmente útiles en la estadística descriptiva, en el análisis exploratorio de los datos.

Pero hay más. Weldon solicitó consejo a Pearson a la hora de analizar las mediciones de cangrejos (diámetro del caparazón, longitud de las patas, etc.) que había realizado durante unas vacaciones en la bahía de Nápoles. Las observaciones no parecían distribuirse de acuerdo a la ley normal. Su distribución no era si-

Representación del gráfico III del artículo «Sobre ciertas variaciones correladas en *Carcinus moenas*», publicado por Weldon en 1893, que recoge la distribución asimétrica —descompuesta en dos curvas normales superpuestas— de las medidas de los cangrejos napolitanos.

FIG. 4



métrica: en lugar de una única montaña, como en la distribución normal, parecían dibujarse dos jorobas (figura 4). Ayudado por Pearson, Weldon diseccionó la distribución en dos componentes normales, siguiendo el pensamiento de Galton de que todas las distribuciones eran normales o mixtura de normales, y concluyó precipitadamente que debía de tratarse de dos especies diferentes de cangrejos que por desconocimiento había medido de modo conjunto o, en su defecto, de una única especie en proceso de generar dos especies diferentes. Pero el matemático inglés quería encontrar una manera de interpretar los datos sin forzar su normalización, sin distorsionar la forma de la curva de frecuencias. No debía descartarse que hubiese una asimetría real en los datos de partida.

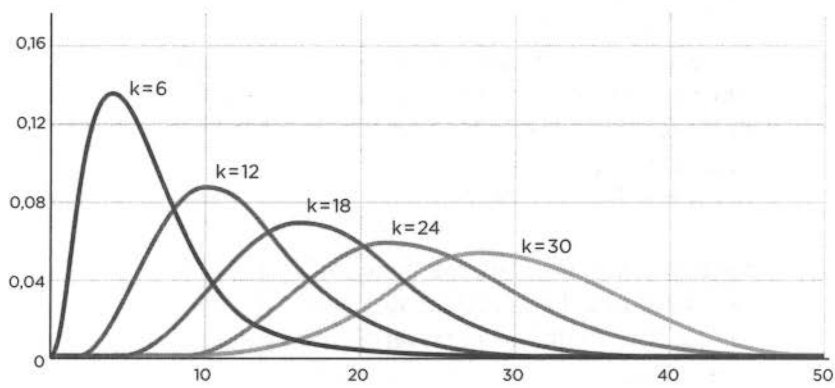
En 1894, en la que sería la primera de sus memorias publicadas sobre estadística, Pearson imaginó todo un sistema de curvas de frecuencias que pudiesen ser de utilidad en las investigaciones biológicas. Quería dotar a los biómetras de un catálogo de modelos que les permitiera extraer toda la información contenida en los datos sin deformarlos. El sistema de curvas de frecuencias permitió disponer, de rebote, de toda una serie de distribuciones de probabilidad que podían aplicarse a distintos fenómenos aleatorios. Entre ellas se cuentan algunas de las distribuciones que más adelante demostrarían ser claves para la extensión de los métodos estadísticos: por ejemplo, la distribución beta, la gamma

o la χ^2 (figura 5). Esta familia de distribuciones asimétricas constituía una alternativa a la distribución normal, dominante desde los tiempos de Quetelet, y lograba mejores ajustes en situaciones prácticas. Para decidir cuál de las curvas había que ajustar a los datos en cada circunstancia, Pearson desarrolló el método de los momentos, que permitía estimar los parámetros que definían cada curva —los llamados *momentos* (un nombre que tomó prestado de la mecánica)— a partir de los datos observados. Este método es el más antiguo conocido para la estimación de parámetros y consiste, en suma, en igualar los momentos apropiados de la distribución teórica con los correspondientes momentos calculados a partir de los datos observados, despejando a continuación los parámetros desconocidos. En concreto, la estimación se realizaba a partir del cálculo de cuatro momentos, relacionados respectivamente con la media, la desviación típica, la asimetría y la curtosis (aunque este término no apareció como tal hasta 1905), que codifican la forma de la curva de frecuencias.

Pearson trataba de desbancar a la distribución normal de su papel preponderante en biología ofreciendo una serie de curvas alternativas para describir distribuciones asimétricas o, incluso, bimodales; porque durante años toda distribución empírica que dibujaba una curva era gaussiana, ya que era todo lo que podía ser. Galton creía ingenuamente que todos los datos tenían que aco-

La distribución χ^2 se trata más bien de una familia de distribuciones, dependiente cada una de ellas de un parámetro denominado «número de grados de libertad» (conforme aumentan, la curva va perdiendo asimetría y converge a una distribución normal).

FIG. 5



modarse a la distribución normal. Pearson, en cambio, enfatizaba que las distribuciones de frecuencias empíricas podían tomar cualquier forma. La curva normal no era la curva canónica, de modo que la tiranía de la ley normal concluyó con el fin de siglo, cuando Pearson consiguió que se aparcara esta visión monolítica. Aparte de la distribución binomial de Bernoulli y de la entronizada distribución normal (ambas relacionadas entre sí por el teorema central del límite), hasta el desembarco del sistema de curvas de Pearson no se disponía de muchos modelos de probabilidad alternativos, con la excepción, entre otras, de la distribución uniforme, la distribución exponencial o la puesta al día de la distribución de Poisson o de los «sucesos raros», popularizada en la época por

LA ALTURA DEL NEANDERTAL

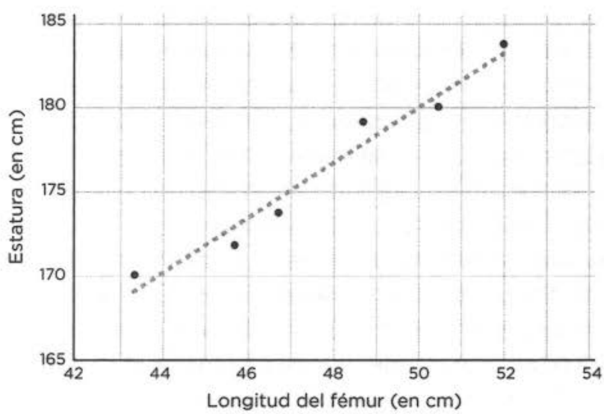
Karl Pearson aplicó el cálculo del coeficiente de correlación y de las rectas de regresión a los datos de las alturas de padres e hijos tomados por Galton. La estatura de los hijos estaba relacionada con la estatura de los padres, de manera que los hijos de padres altos solían ser altos. No había una relación matemática perfecta, pero existía una tendencia, que podía medirse mediante el «coeficiente de correlación de Pearson» (que se define como el cociente entre el momento-producto o covarianza y las desviaciones típicas de las dos variables bajo estudio). Los valores de este coeficiente siempre estaban entre –1 y +1. Si el coeficiente de correlación estaba cerca de 1 significaba que cuando la variable «estatura del padre» aumentaba, la variable «estatura del hijo» también lo hacía. En 1898 Pearson conjeturó que un comportamiento similar se daba entre la estatura de un hombre y la longitud de su fémur. Estudiando cientos de mediciones, encontró que la correlación entre la estatura y la longitud del fémur era de 0,8048. Se trataba de una correlación directa fuerte. A continuación, dedujo la relación existente entre la longitud del fémur y la estatura total del individuo. En otras palabras, determinó la recta de regresión de la estatura sobre la longitud del fémur, hallando en el caso de los varones:

$$\text{Estatura (en cm)} = 81{,}31 \text{ cm} + 1{,}88 \cdot \text{Longitud del fémur (en cm)}.$$

Finalmente, Pearson enseñó cómo usarla para reconstruir la estatura de los hombres prehistóricos a partir de las medidas de sus huesos. Por ejemplo, en el caso del hombre de Neandertal, la longitud media del fémur era de 44,52 cm, con lo que sustituyendo en la ecuación de arriba se obtenía que su

representar el porcentaje de oficiales prusianos que en la década de 1890 resultaron heridos por las coces de sus caballos.

Aún más, en 1896, Pearson logró la definitiva matematización del coeficiente de correlación y de la regresión lineal, que Galton manejara empíricamente. Mientras que Galton empleaba unas matemáticas muy modestas y raramente trabajaba con más de 100 datos (para así usar porcentajes cómodamente), Pearson hizo de la matemática abstracta un requisito para hacer estadística y tomó en consideración grandes conjuntos de datos (más de 1 000). Ofreció tanto la fórmula del coeficiente de correlación en que aparece el «momento-producto» (lo que Fisher y su círculo llamarían *covarianza*, un nombre que ha hecho fortuna) como las ecuaciones ex-



Este diagrama de dispersión relaciona la longitud del fémur y la talla (en centímetros) de una muestra de seis individuos. Como puede observarse, entre ambas variables existe una correlación lineal fuerte (con línea punteada, la recta de regresión).

estatura promedio era de 165,01 cm. Por su parte, el hombre de Cro-Magnon medía 172,15 cm, dado que la longitud media de los fémures conservados era de 48,32 cm. Tanto el hombre de Neandertal como el de Cro-Magnon eran sensiblemente más bajos que los hombres actuales. En esencia, esta es la metodología que a día de hoy siguen empleando los paleoantropólogos para inferir las características de las especies extintas de homínidos que desentierran en las excavaciones.

plícitas de las rectas de regresión, aunque no completó la teoría de la regresión no lineal (curvilínea) hasta 1905. Su ayudante en aquel tiempo, el ingeniero y luego profesor de Estadística George Udny Yule, desarrolló hacia 1897 la regresión múltiple (en más de dos variables, cuando se supone que la variable de estudio depende de dos o más), conectándola con el método de mínimos cuadrados y la síntesis Gauss-Laplace. Es poco conocido que Pearson fue el primero en alertar del peligro de la detección de «correlaciones espurias» (uno de los abusos que cometería con la estadística la segunda mitad del siglo xx): dos variables pueden estar fuertemente correlacionadas entre sí sin que entre ambas medie una relación de causa-efecto o ni siquiera una causa común (como es el caso, por ejemplo, del número de cigüeñas presentes en Londres y el número de niños nacidos cada semana en esa ciudad).

Finalmente, en 1900, Pearson publicó el test de la chi-cuadrado (χ^2) para comprobar la bondad del ajuste entre la distribución observada y la distribución teórica o esperada. El test demostró ser útil no solo para dar una medida del ajuste entre datos y distribuciones, sino que fue generalizado por Pearson y sus discípulos para contrastar la homogeneidad entre varias muestras y la independencia entre variables (aunque el número exacto de grados de libertad de la distribución χ^2 que interviene en el test lo facilitó Fisher en la década de 1920). En consecuencia, la adjudicación de una distribución normal ya no era cuestión de una semejanza percibida cualitativamente entre gráficas, sino de una significación estadística cuantitativa. Se trataba de uno de los puentes más sólidos tendidos hasta el momento entre la estadística descriptiva y la estadística inferencial. De hecho, a finales del siglo xx una conocida revista científica estadounidense eligió el test χ^2 como uno de los veinte descubrimientos científicos del siglo que más había cambiado nuestras vidas.

Entre otras innovaciones más prosaicas, Pearson y sus colaboradores publicaron toda una serie de tablas para biómetras y estadísticos de gran ayuda en el ajuste de curvas, y para cuyo diseño se sirvieron de máquinas de calcular pioneras. No hay que olvidar que hasta el advenimiento del ordenador, estas tablas simplificaban enormemente la vida a los estadísticos, permitién-

doles consultar de un vistazo el resultado de laboriosos cálculos de probabilidades. Esta abundante cosecha de resultados fue dada a conocer a lo largo de un total de dieciocho artículos que Pearson escribió entre 1894 y 1912 bajo el título común de *Contribuciones matemáticas a la teoría de la evolución*. Hoy día estos artículos son un claro indicador de la extraordinaria capacidad para trabajar y relacionar materias dispares de que hacía gala Karl Pearson.

LA INSTITUCIONALIZACIÓN DE LA ESTADÍSTICA

Los primeros artículos de Pearson vieron la luz dentro de las *Philosophical Transactions* de la Royal Society, pero la oposición despertada entre los biólogos de la sociedad por los prolijos análisis matemáticos de los datos (los naturalistas no estaban dispuestos a aceptar conclusiones biológicas sobre la base de razonamientos estadísticos) condujo a Weldon y a Pearson a fundar, con el apoyo de Francis Galton, la revista *Biometrika* en 1901. La idea de crear una revista propia para publicar las investigaciones se debió a Weldon, pero fue Pearson quien sugirió su peculiar nombre. Para ambos científicos, el problema de la evolución era un problema estadístico. Darwin había planteado su teoría biológica sin recurrir a la matemática, pero cada uno de sus conceptos, desde la variación y la selección a la herencia y la regresión, era susceptible de ser definido matemáticamente y analizado estadísticamente.

En el editorial de presentación de la revista, Weldon y Pearson describían su radio de acción y profetizaban el advenimiento de un día en que habría matemáticos que serían competentes biólogos y, recíprocamente, biólogos que serían competentes matemáticos. Durante varios lustros, *Biometrika* publicó sesudos análisis estadísticos sobre datos tan dispares como la envergadura de los pájaros exóticos, la altura de los reclutas albaneses, la medida de la tibia de los nativos africanos o la longitud del pene de los pigmeos.

LA χ^2 Y LOS V2 DISPARADOS POR LOS NAZIS CONTRA INGLATERRA

Durante la Segunda Guerra Mundial los alemanes lanzaron una lluvia de cohetes V2 sobre Londres. Los estadísticos que colaboraban en la defensa antiaérea dividieron el mapa de Londres en cuadrículas de 1/4 km² (hasta un total de 576) y contaron el número de bombas caídas en cada cuadrícula durante un bombardeo alemán. Observaron que en 229 cuadrículas no caía ninguna bomba; en 211 caía solo una, etcétera. Los resultados fueron:

Nº de impactos en la cuadrícula	0	1	2	3	4	5
Frecuencia observada	229	211	93	35	7	1

Los estadísticos querían averiguar si los bombardeos seguían un patrón aleatorio, es decir, si no estaban dirigidos a determinados objetivos militares, de manera que el vuelo de los V2 estaba todavía lejos del control de los científicos alemanes. Para ello emplearon el test χ^2 de Pearson, con el propósito de comprobar el ajuste entre la distribución observada y la distribución teórica esperada, que en este caso se trataba de una distribución de Poisson o de los «sucesos raros», ya que esta última mide la probabilidad de que aleatoriamente ocurra un determinado número de eventos —que se suponen «raros», improbables— durante cierto período de tiempo. La distribución de Poisson depende únicamente de un parámetro, habitualmente denotado como λ, que representa la frecuencia de ocurrencia media. El valor estimado de λ a partir de los datos empíricos es:

$$\lambda = \frac{0 \cdot 229 + 1 \cdot 211 + 2 \cdot 93 + \ldots + 5 \cdot 1}{576} = 0{,}929$$

(en promedio, uno esperaría aproximadamente un impacto por cuadrícula). En consecuencia, las frecuencias que debían esperarse si los bombardeos se ajustaban a esta distribución eran las siguientes (la fórmula de donde salen estos valores es un poco aparatosa pero fácil de justificar, aunque aquí no entraremos en ello):

	Número de impactos/cuadrícula					
	0	1	2	3	4	5
Frecuencia esperada	227,5	211	98	30	7	1,5
Frecuencia observada	229	211	93	35	7	1
Discrepancias	1,5	0	-5	5	0	-0,5

A continuación, los estadísticos determinaron el valor del «estadístico chi-cuadrado», que es una medida de la discrepancia total que se calcula sumando las diferencias entre la frecuencia observada y la frecuencia esperada elevadas al cuadrado (así no se compensan las discrepancias positivas con las negativas) y dividiendo por la frecuencia esperada:

$$\chi^2 = \sum \frac{(\text{Discrepancia})^2}{\text{Frecuencia esperada}} = \frac{1{,}5^2}{227{,}5} + \ldots + \frac{(-0{,}5)^2}{1{,}5} = 1{,}27.$$

Si la distribución de Poisson era la adecuada, este estadístico era un valor de una distribución chi-cuadrado con 6 - 2 = 4 grados de libertad (en general, es siempre uno menos que el número de clases de partida, pero como hemos estimado el valor de λ a partir de los datos, hay que restar uno más según demostró Fisher). Consultando las tablas, los estadísticos observaron que la probabilidad de que una χ^2_4 tome un valor mayor o igual que 1,27 es de 0,87. En otras palabras, la probabilidad de obtener una discrepancia como la observada era significativamente alta bajo el supuesto de que los bombardeos se producían aleatoriamente, sin un objetivo fijo. Los londinenses podían respirar tranquilos.



Misil V2 en su plataforma.

Karl Pearson fue editor continuado de la revista *Biometrika* desde su primer número, publicado en octubre de 1901, hasta su muerte, ocurrida treinta y cinco años después. Tras el inesperado fallecimiento de Raphael Weldon en un desafortunado accidente de esquí en 1906, Pearson se alejó de la biología evolutiva. Sin la inestimable colaboración de su bien entrenada mente biológica, Pearson no se sentía con fuerzas para proseguir en solitario con el estudio estadístico de la evolución y la herencia. Sin embargo, redobló esfuerzos en la institución de un centro que convirtiera la estadística en una rama de la matemática aplicada con vida propia, con una nomenclatura y unos métodos independientes, de manera que los estadísticos fuesen por derecho propio «hombres de ciencia».

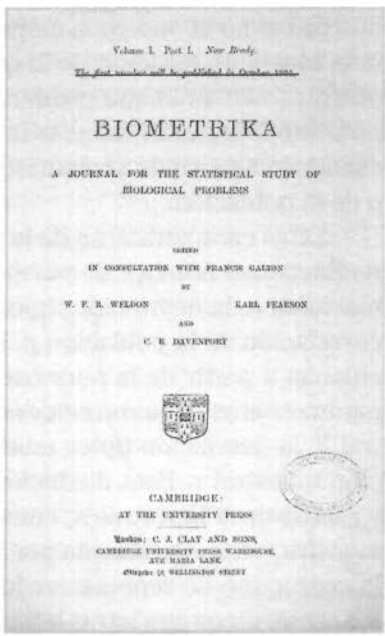
«La ciencia del futuro se llamará biometría y su órgano oficial será *Biometrika.*»

— Karl Pearson.

Sir Francis Galton falleció en 1911, dejando en herencia la provisión de una cátedra de Eugenesia en el University College de Londres, que fue ocupada por su protegido, Pearson, quien hizo así realidad su sueño de formar un Departamento de Estadística Aplicada combinando el Laboratorio Biométrico (que dirigía desde su fundación en 1903) y el Laboratorio Galton para la Eugenesia Nacional (surgido en 1907 como evolución de la *Eugenics Record Office*, instituida por Galton en 1904). El Laboratorio Biométrico desarrollaba los métodos estadísticos en un contexto biológico, mientras que el Laboratorio Eugenésico los aplicaba en el estudio del «deterioro nacional» (relacionando, por ejemplo, las tasas de fertilidad con el estatus social o el alcoholismo con su influjo en el físico y la habilidad de la descendencia). En 1925, coincidiendo con la especialización de *Biometrika* en temas estadísticos teóricos, Pearson fundó *Annals of Eugenics* (actualmente rebautizada como *Annals of Human Genetics*), para proseguir con la publicación de investigaciones prácticas sobre la eugenesia.







Volume I. Part I. *Now Ready.*

~~*The first number will be published in October 1901.*~~

BIOMETRIKA

A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS

EDITED

IN CONSULTATION WITH FRANCIS GALTON

BY

W. F. R. WELDON KARL PEARSON

AND

C. B. DAVENPORT

CAMBRIDGE:
AT THE UNIVERSITY PRESS

London: C. J. CLAY AND SONS,
CAMBRIDGE UNIVERSITY PRESS WAREHOUSE,
AVE MARIA LANE.
Glasgow: 50, WELLINGTON STREET

FOTO SUPERIOR:
Fotografía tomada en 1909 que muestra a un anciano Galton de ochenta y siete años acompañado por Karl Pearson.

FOTO INFERIOR IZQUIERDA:
Karl Pearson con un busto de Raphael Weldon. La fotografía es de 1910.

FOTO INFERIOR DERECHA:
Cabecera original de *Biometrika*, la revista editada por Weldon y Pearson con el apoyo de Galton y la colaboración de Charles Davenport (1866-1944), prominente biólogo estadounidense que compartía el enfoque biométrico y el credo eugenésico.

UNA POLÉMICA ENCARNIZADA

En 1914 Pearson recibió un artículo firmado por un profesor de escuela de veinticuatro años llamado R.A. Fisher para ser publicado en la revista que dirigía y editaba, *Biometrika*. En las apretadas páginas del borrador, Fisher deducía un resultado que a Pearson y su equipo se les había escapado sistemáticamente: la distribución correcta del coeficiente de correlación muestral r, un conocimiento necesario para determinar el error probable a la hora de estimar el coeficiente de correlación poblacional ρ. La cuestión de las distribuciones en el muestreo había comenzado a percibirse como un tema candente para el progreso de la inferencia estadística, por cuanto permitía cuantificar la fiabilidad de las predicciones realizadas en base a una muestra representativa con el fin de conocer determinadas características de una población, de una colectividad que se presupone demasiado numerosa como para ser estudiada exhaustivamente. Proporcionar una estimación de la correlación ρ en toda la población a partir de la correlación r observada en los datos de la muestra era engañoso y de escasa utilidad si no se indicaba su precisión. El estudio de la distribución muestral, es decir, de la que resulta de considerar todas las posibles muestras que pueden extraerse aleatoriamente de una población, permitía calcular la probabilidad de que el valor de r calculado a partir de una muestra se acerque al valor desconocido ρ de la población.

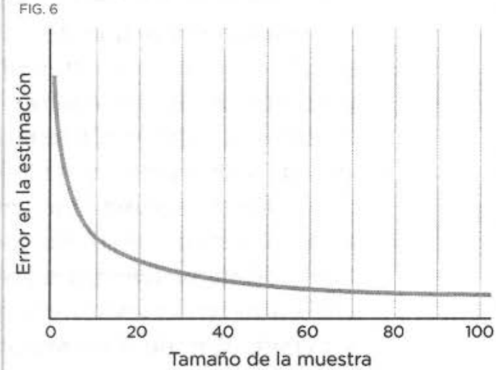
Estas características de la población que se deseaba estimar recibieron el nombre de *parámetros*. Por ejemplo: la media poblacional μ, la desviación típica poblacional σ o el coeficiente de correlación de la población ρ. En cambio, los valores que se calculaban a partir de la muestra para estimar puntualmente estos parámetros se llamaron *estadísticos*. Por ejemplo: la media muestral $\bar{X}$, la desviación típica muestral S o el coeficiente de correlación muestral r. Esta distinción entre parámetros poblacionales y estadísticos muestrales, como la subyacente entre población y muestra, sería canonizada por Fisher. Aún más: a Fisher se debe la costumbre de representar los parámetros con letras griegas y los estadísticos con letras latinas, con la excepción de la notación

barrada para la media muestral, que deriva de Maxwell. Ahora bien, junto a la estimación, se deseaba dar un valor de la variación o dispersión de todas las posibles estimaciones, a fin de dar una idea de la exactitud de la inferencia. Para ello se calculaba el error probable o, también, el «error estándar» del estimador, que no es más que la desviación típica de la distribución del estadístico en el muestreo (esto es, de la distribución que mide la probabilidad de que el estadístico tome tal o cual valor en función de los datos de la muestra, que se considera que han sido seleccionados aleatoriamente de la población). Este número decía lo buena que era la inferencia: a menor error, mejor estimación. Además, el error suele depender de la raíz cuadrada del tamaño de la muestra, de manera que conforme el tamaño muestral aumenta, la precisión de la estimación también lo hace, ya que el error disminuye con la raíz cuadrada del tamaño (figura 6).

Años antes, en 1896, Pearson había enunciado, sin demostrarlo —la demostración corría a cargo de Fisher—, que el estimador más probable de ρ, de la correlación de toda la población, era en esencia r, la correlación calculada a partir de los datos observados en la muestra (aunque la notación de Pearson no distinguía bien entre ambos valores, entre el parámetro poblacional y el estadístico muestral). Pearson respondió con entusiasmo a Fisher, felicitándole por la prueba y transmitiéndole que el artículo sería sin duda aceptado. Una semana después, Pearson volvía a escribir a Fisher, contándole que por fin había leído con detenimiento el borrador, que le parecía que era un avance y que sería un honor publicarlo si ampliaba un poco las páginas del final. Fisher estudiaba la distribución del coeficiente de correlación muestral geométricamente (imaginando la muestra como un vector n-dimensional y la distribución como una variedad diferenciable) y recurría, además, a una transfor-

El error cometido en la estimación disminuye rápidamente con el tamaño de la muestra, hasta un punto en el que el aumento del tamaño muestral no se traduce en una reducción apreciable del error.

FIG. 6



mación algebraica, con lo que a Pearson le costaba seguir una demostración en que no se razonaba a partir de r sino de una función definida sobre r. Fisher dio la bienvenida a la sugerencia y su artículo revisado fue felizmente publicado en *Biometrika* en 1915.

Hasta 1917 la relación entre ambos matemáticos fue cordial, pero en la primavera de ese año Pearson y sus colaboradores publicaron un estudio cooperativo, en el que Pearson arremetía contra Fisher, dedicando más de una página a criticar un supuesto error cometido por este último en su artículo de 1915. Quizá obró así movido por la nota que Fisher le había enviado cuestionando la investigación llevada a cabo por una doctoranda danesa que trabajaba en el laboratorio de Pearson; además, parecía poner en duda los méritos del test χ^2 y del método de los momentos para construir estimadores. En el artículo mencionado de 1915, Fisher daba cumplida demostración de la afirmación que Pearson hiciera bastantes años antes: el valor más probable del coeficiente de correlación ρ de toda una población es, en esencia, el coeficiente de correlación r observado en la muestra (cuando el tamaño muestral crece, porque en general r tiende a ser mayor que ρ). Pearson afirmaba que Fisher lo había demostrado empleando los métodos inversos de probabilidad, es decir, el teorema de Bayes, ocasión que aprovechó para dirigirle una reprimenda, señalando lo arbitrario del procedimiento, ya que tenía que partir de una distribución *a priori* uniforme, de una presuposición de ignorancia total. Sin embargo, Fisher no había usado este procedimiento. Como ampliaremos en el capítulo 5, Fisher no solo compartía esta oposición radical a la inferencia bayesiana, sino que había empleado otro método, un método nuevo que explicaremos en el próximo capítulo: el «método de máxima verosimilitud», que poco o nada tenía que ver, pero que ciertamente venía expresado con términos ambiguos.

A Fisher no tuvo que agradarle la lectura de este pasaje del estudio, y es lógico que el incidente le pesara a la hora de declinar la oferta de trabajar a las órdenes de Pearson en el Laboratorio Galton y decantarse por ocupar la plaza de estadístico en la Estación Agrícola Experimental de Rothamsted a partir de 1919. Además, Fisher elaboró una respuesta en forma de artículo que le

hizo llegar a Pearson en 1920. Allí profundizaba en el estudio del coeficiente de correlación para una muestra pequeña y, de paso, indicaba que en su artículo de 1915 no había empleado para nada el teorema de Bayes. Y aunque decía mostrarse reacio a criticar a los estadísticos autores del estudio (entre ellos, claro está, Pearson), llegaba al extremo de ridiculizar los ejemplos que ponían, terminando su respuesta con una nota sobre la confusión entre la regla de Bayes y su nuevo método de construcción de estimadores. Como es natural, Pearson rechazó tajantemente publicar el artículo y se lo devolvió a su autor, rogándole que no insistiera.

El principal resultado de esta desafortunada controversia fue una enemistad declarada que se prolongó durante años, de manera que ninguno de los dos estadísticos desaprovechaba la ocasión de poder criticar al rival. Tanto es así que cuando Fisher, en una trilogía de artículos publicados entre 1922 y 1924, perfeccionó el test de la chi-cuadrado, dando el número exacto de grados de libertad, Pearson nunca aceptó la modificación, pese a ser correcta. Recíprocamente, cuando en 1945 se solicitó a Fisher que escribiera la entrada sobre Pearson para un diccionario de biografías, el editor hubo de rechazar de plano su texto por el tono calumnioso que emanaba. En cualquier caso, soslayando las rencillas académicas, hay que poner de relieve el acusado contraste entre las visiones de la estadística de Pearson y Fisher, por cuanto el primero empleaba muestras grandes y el segundo, por el contrario, influido por William Sealy Gosset (alias *Student*), prefería trabajar con muestras pequeñas, amparándose en el dicho estadístico que afirma que para catar la sopa, aunque la olla sea más grande, basta con una cucharada pequeña.

Karl Pearson jugó un papel enorme en determinar el contenido y la organización de la investigación estadística en su día, a través de sus investigaciones, sus enseñanzas, el establecimiento de laboratorios y el inicio de un vasto programa de publicaciones. A una obra tan prolífica que no tiene rival en cantidad en ningún otro matemático, hay que añadir una capacidad de trabajo inmensa, que el propio Pearson achacaba, con una pizca de ironía, a que nunca contestaba al teléfono ni asistía a comités de bienvenida.

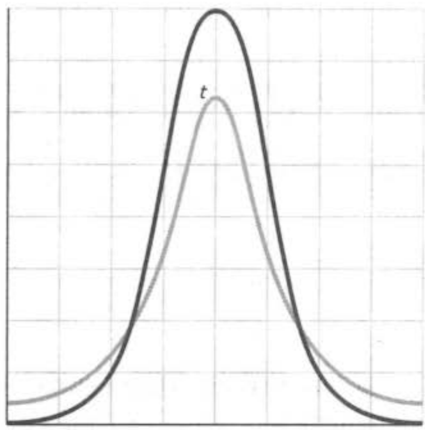
STUDENT Y LA DESTILERÍA GUINNESS

William Sealy Gosset (1876-1937) era químico de formación, aunque se había familiarizado con la estadística tras pasar una temporada en el Laboratorio Biométrico con Pearson. En 1908 publicó un célebre artículo, titulado «El error probable de la media», bajo el seudónimo *Student*. La razón es que la empresa para la que trabajaba, la fábrica de cerveza Guinness en Dublín, no permitía que los empleados hicieran públicas las investigaciones que realizaban para la marca. Buscando controlar la calidad de la cerveza producida, Student recogía muestras pequeñas (lo que salía más barato). Y había descubierto que uno de los tipos de curvas de Pearson era una distribución de probabilidad de gran utilidad para el estudio de estos experimentos a pequeña escala. Si, por ejemplo, quería estimar la acidez media de toda la cerveza producida por la planta en un cierto período de tiempo, calculaba la media de los niveles de acidez encontrados en la docena de barriles de muestra. El problema, y de ahí el título del artículo, es que Student no conocía el error probable que cometía en la estimación de la media poblacional por medio de la media muestral, un número necesario para valorar si la inferencia era o no precisa y, dicho sea de paso, si la acidez entraba dentro de los límites aceptables. Para determinarlo, Student precisaba conocer la distribución de probabilidad del estadístico media muestral. Se sabía que si la muestra era grande —en la práctica, mayor o igual que 30—, la distribución de la media muestral era normal (en virtud del teorema central del límite). Pero si la muestra era pequeña, no tenía por qué serlo.

La distribución *t* de Student
Student obtuvo la distribución correcta, conocida hoy día —después de que Fisher la retocara en 1925— como distribución *t* de Student. Esta distribución es, en realidad, una familia de distribuciones dependientes del número de grados de libertad; pero, en general, es más aplanada que la distribución normal, con colas más anchas, lo que refleja la mayor incertidumbre de las inferencias. Este modelo de probabilidad es imprescindible en el presente por su robustez, ya que no solo se emplea en la inferencia a partir de muestras pequeñas extraídas de una población normal (de la que se desconocen su media y su desviación típica), sino también cuando la población subya-

Todas las piezas del rompecabezas estaban ya sobre la mesa. Todo estaba listo para el reordenamiento de los materiales estadísticos que iba a realizar Fisher. De resultas, la estadística sería encumbrada como un nuevo estilo de razonamiento, que se sumaría, en el plano teórico, al razonamiento axiomático matemático y, en el plano experimental, tanto al método hipotético-deductivo

cente no se distribuye normalmente. La distribución *t* es prácticamente insensible al supuesto de normalidad.



Como puede observarse, la *t* de Student (en gris) presenta colas más anchas que la normal (en negro).

Rescatado del olvido

No obstante, Student fue una figura marginal hasta que Fisher rescató su labor del olvido, aunque estaba dotada de un sentido del humor peculiar (como se observa en la regla mnemotécnica que inventó en relación con la curtosis: para recordar el término «platicúrtico», que se aplica a las curvas más aplanadas que la normal, Student se acordaba de un *platypus*, ornitorrinco en español; y para recordar el término «leptocúrtico», aplicable a las curvas más puntiagudas, traía a la memoria un par de canguros entrechocando sus cabezas, porque *lepping* significa saltando en inglés). Fisher y Student establecieron contacto alrededor de 1912, por mediación del tutor del primero en Cambridge, un astrónomo de reconocido prestigio. Los apuros que Student mostraba por carta con las demostraciones matemáticas inspiraron a Fisher la posibilidad de deducir exactamente la distribución de varios estadísticos en el muestreo y, de este modo, anotarse sus primeros éxitos. Por su parte, la apatía de Pearson al respecto se explica porque estaba convencido de que la detección de las pequeñas tendencias que se observaban en los datos biológicos requería del empleo de muestras grandes, de un gran número de datos: «¡Solo los sucios cerveceros manejan muestras pequeñas!», solía decir con tono jocoso a sus ayudantes.

de la física como al taxonómico de las ciencias naturales. La estadística se convertiría en un nuevo modo de pensar y, en especial, de hacer, de intervenir en el mundo, aplicándose en áreas tan dispares como los laboratorios biométricos, las granjas agrícolas o la industria cervecera. Una nueva estrella anunciaba su salida en el firmamento.

CAPÍTULO 3

Los fundamentos matemáticos de la inferencia estadística

En los años veinte, Fisher tomó el relevo de la primera generación de estadísticos, crecida en torno a Pearson. Su artículo «Sobre los fundamentos matemáticos de la estadística teórica» fue el aldabonazo que anunció la implantación de la inferencia estadística como disciplina matemática, seguido por dos influyentes libros: *Métodos estadísticos para investigadores* y *El diseño de experimentos*. En ellos, Fisher cimentaría los test de significación, el análisis de la varianza y la aleatorización como principios básicos de cualquier confrontación del científico natural con los hechos.

La inferencia estadística se define como una colección de técnicas que permiten formular inferencias de lo particular (la muestra) a lo general (la población), proporcionando —y esto es lo que separa a la estadística de la adivinación— una medida de la incertidumbre de la predicción: la probabilidad de error.

Según se ha visto en los capítulos anteriores, la unión entre los rudimentarios métodos estadísticos de Laplace y Gauss, confinados al espacio de la astronomía, y la ciencia del Estado, circunscrita al campo de la demografía y la incipiente ciencia social, se produjo a caballo entre los siglos XIX y XX en el terreno intermedio de la biología, ya que la evolución se reformuló como problema estadístico gracias al influjo de la eugenesia y la biometría.

La estadística prefisheriana, dominada por ese titán llamado Karl Pearson, se encontraba en la siguiente situación. En estadística descriptiva, aunque no se distinguía claramente entre población y muestra, se conocían las representaciones gráficas más comunes (diagrama de barras, histograma, diagrama de dispersión, etc.) y se calculaban las principales medidas de centralización (media, mediana, moda), dispersión (la desviación típica, aunque no era la única medida), posición (cuartiles y percentiles) y forma (asimetría y curtosis). El viaje desde el análisis exploratorio de los datos al dominio de la teoría matemática de la probabilidad se realizaba mediante el ajuste de distribuciones teóricas —la curva

normal o las curvas de Pearson— sobre las distribuciones de frecuencias observadas, por medio del método de mínimos cuadrados y del método de los momentos. La bondad del ajuste podía comprobarse mediante ese gran invento que era el test de la χ^2. Finalmente, el establecimiento de inferencias estadísticas solo contaba con dos métodos expeditos: las predicciones fundadas en el análisis de la regresión y la correlación; y, en especial, los métodos inversos de probabilidad, mayoritarios desde el tiempo de Laplace y basados en el teorema de Bayes (la inferencia bayesiana o subjetiva).

Fisher vendría a rellenar el vacío de este importantísimo cajón planteando gran parte de los métodos de estimación e inferencia hoy clásicos (la inferencia frecuentista u objetiva). Si Pearson había enseñado cómo extraer información relevante de la maraña de datos, Fisher mostró cómo conocer el todo (la población) observando la parte (la muestra). Él fue el arquitecto que afianzó definitivamente el puente entre la estadística descriptiva y la estadística inferencial, atando esta última a la matemática, lo que insufló nuevos aires a la disciplina.

Y lo que es más importante, Fisher estructuró las etapas del método estadístico. Al análisis exploratorio inicial de los datos disponibles y la construcción de un modelo probabilístico tentativo, le seguiría una fase de estimación de los parámetros desconocidos del modelo poblacional a partir de la muestra observada y, finalmente, otra fase de ajuste entre el modelo y la realidad por medio de los test de significación y el diseño de experimentos. Con Fisher puede decirse que culminó el cierre del cuerpo metodológico básico de la estadística: la elección del modelo teórico a partir de los datos empíricos, la deducción matemática de las propiedades del mismo, la estimación de los parámetros desconocidos y la validación final del modelo mediante un test experimental. Esta aproximación, en la que se recoge información de los resultados de un experimento y a partir de ellos se intenta sacar conclusiones, es el núcleo de la inferencia estadística, que a diferencia del cálculo de probabilidades no es un razonamiento deductivo sino inductivo, sometido a cierto error que se busca cuantificar.

PROBLEMAS Y CRITERIOS DE LA INFERENCIA

En 1919, Fisher aceptó un puesto como asesor estadístico en la Estación Agrícola Experimental de Rothamsted, tras rechazar la oferta de trabajo de Karl Pearson en el Laboratorio Galton para no tener que sufrir su supervisión, ya que las diferencias entre ambos estaban lejos de limarse. Con veintinueve años se trasladó, junto con su esposa e hijos, a vivir a una vieja granja al norte de Londres, cercana a la estación. Los dueños, fabricantes de abonos, le habían contratado con la intención de que pusiera orden en la enorme cantidad de datos que se habían ido recopilando durante los años de funcionamiento del centro. El tiempo demostraría que la decisión tomada fue la acertada. Sir Edward John Russell (1872-1965), responsable de la estación, mantenía una atmósfera de libertad que estimulaba el intercambio científico entre biólogos, químicos y estadísticos. Fisher se convirtió en un investigador agrario infatigable, y entre la granja y la estación germinaron sus ideas más geniales, sin dejar de lado ninguna parcela de la estadística.

En su artículo seminal titulado «Sobre los fundamentos matemáticos de la estadística teórica» (leído en la Royal Society de Londres en 1921 y publicado en 1922), Fisher acuñó la nomenclatura hoy habitual en cualquier manual de inferencia estadística. Por ejemplo: el término *parámetro*, en su sentido estadístico moderno, aparece por vez primera y se menciona hasta 57 veces. Una afirmación errónea que hiciera el astrofísico Arthur S. Eddington en su libro *Movimientos estelares* (1914), junto a varias preguntas formuladas epistolarmente por Pearson antes de que cortaran el contacto, fueron el punto de partida que espoleó a Fisher para estudiar la cuestión de la estimación estadística en detalle.

Este artículo fundacional arranca señalando que el objeto de los métodos estadísticos es la «reducción» de los datos: expresar toda la información relevante contenida en la muestra sobre la población por medio de unos pocos valores numéricos. Inmediatamente después, Fisher ponía de relieve la noción de «modelo estadístico», que posibilitaba distinguir con claridad entre una población (real o hipotética) y una muestra suya, un par de conceptos conjugados cuya raya de separación había estado hasta el momento difuminada. Los

datos disponibles han de considerarse como una muestra aleatoria proveniente de una población, cuya distribución con respecto a la característica bajo estudio viene especificada por una lista de parámetros que se denotan con letras griegas (por ejemplo, el parámetro θ). En verdad, para cada posible valor de los parámetros, se tiene una población distinta, de modo que la pregunta central que se formula cada estadístico es, a saber: ¿a cuál de las infinitas poblaciones posibles pertenece esta muestra que tengo delante?

A continuación, Fisher indicó las tres clases de problemas matemáticos a que se enfrenta la inferencia estadística. En primer lugar, los problemas de «especificación», que consisten en definir el modelo poblacional, es decir, la familia de distribuciones dependientes de uno o más parámetros θ de la que se extraen (supuestamente) las muestras. En segundo lugar, los problemas de «estimación», que por ser el eje principal de la inferencia estadística se explican más adelante en detalle. Y en tercer y último lugar, los problemas de «distribución», cuyo propósito es deducir exactamente la distribución de un estadístico en el muestreo a partir de la distribución de la población, que se supone conocida. Las distribuciones muestrales determinan la probabilidad con que cierto estadístico toma valores entre dos límites prefijados (equivalentemente, la frecuencia relativa con que los toma cuando el proceso de muestreo se repite indefinidamente). La resolución de esta clase de problemas es, en cierto modo, un requisito previo a la inferencia, pues permite hallar el error estándar cometido en la estimación, así como comparar los méritos de varios estimadores entre sí. Posibilita, en suma, calcular la precisión del estimador y medir la incertidumbre en la predicción del parámetro o parámetros desconocidos.

Centrándonos en los problemas de la teoría de la estimación, Fisher apuntó que se trata de la elección del valor del parámetro θ más apropiado basándose en la muestra o, más exactamente, en los estadísticos —denotados con letras latinas (como, por ejemplo, T)— que se calculan a partir de los datos observados. ¿Por qué se usaba la media muestral $\bar{X}$ para estimar la media poblacional μ? ¿O la desviación típica muestral S para aproximar la desviación típica poblacional σ? La teoría de la estimación estadística

que esbozó Fisher respondió a estas preguntas planteándose qué propiedades debía cumplir todo buen estimador.

Una primera propiedad que parecía natural exigir a la hora de estimar un parámetro θ mediante un estadístico T era que fuera «consistente», es decir, que T convergiera en probabilidad al verdadero valor de θ conforme el tamaño de la muestra aumentara. En consecuencia, si la muestra era grande, el valor de T coincidiría muy probablemente con el de θ. Los estimadores consistentes eran aquellos que se volvían mejores según crecía el tamaño de la muestra, que tendían a dar el valor correcto del parámetro.

> «Hay que admitir que cualquier inferencia de lo particular a lo general se realiza con un cierto grado de incertidumbre, pero esto no es lo mismo que admitir que esa inferencia no pueda ser absolutamente rigurosa.»
>
> — R.A. FISHER, *EL DISEÑO DE EXPERIMENTOS* (1935).

Una segunda propiedad deseable era que T no solo convergiera al valor real del parámetro θ, sino que lo hiciera de manera «eficiente», es decir, con el menor error posible. En términos más precisos: que el error estándar del estimador fuera el mínimo posible (más adelante veremos que Fisher dio con un método —el método de máxima verosimilitud— para construir estimadores eficientes).

Finalmente, una tercera condición, más restrictiva que la de eficiencia, era la propiedad de «suficiencia», que pedía que el estadístico T no desaprovechara ninguna información contenida en la muestra, que contuviera toda la información relevante para estimar el parámetro correspondiente. Cuando un estadístico T era suficiente para θ, ningún otro estimador proporcionaba más información sobre el parámetro a partir de la muestra. Además, podía demostrarse que en este caso T era eficiente. La suficiencia era el criterio supremo, ya que implicaba los otros dos criterios más débiles (la eficiencia y la consistencia). Cuando se encontraba un estadístico suficiente, podía afirmarse que el problema de la estimación estaba completamente resuelto. Por desgracia, no

SESGO Y EFICIENCIA

A día de hoy, los tres criterios proporcionados por Fisher apenas han experimentado modificación, aunque su acción se ha visto complementada por otros criterios.

El sesgo
Así, se comienza definiendo un estimador *T* como centrado o insesgado para el parámetro θ si, para cualquier tamaño muestral, la media de su distribución en el muestreo es θ. En otras palabras, si el valor esperado del estadístico *T* es, precisamente, el valor real de θ. En otro caso, se dice que el estimador no es centrado, que tiene sesgo.

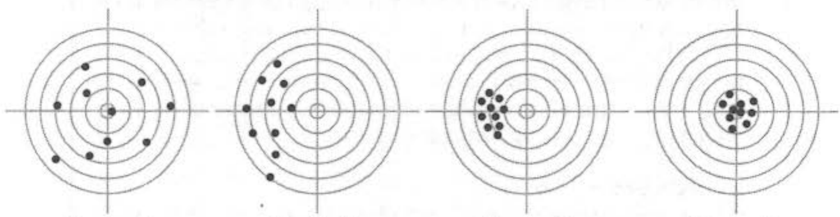
La eficiencia
Por su parte, se llama *eficiencia* o *precisión* de un estimador al inverso de la varianza de su distribución muestral, es decir, al inverso del cuadrado de su desviación típica, de su error estándar (el concepto de *varianza* como cuadrado de la desviación típica fue introducido por Fisher en 1918 por ser más cómodo de calcular). La eficiencia o precisión de un estimador está, por tanto, ligada a su varianza (ambas cantidades son inversamente proporcionales entre sí): cuanta más dispersión tiene un estimador, menos preciso es en sus estimaciones, y recíprocamente. Este concepto es especialmente relevante para comparar estimadores insesgados, ya que entre ellos es preferible el más eficiente, el de mínima varianza.

El error cuadrático medio
No obstante, en ocasiones se presenta el dilema de elegir entre dos estimadores con propiedades contrapuestas: uno de ellos es insesgado, mientras que el otro es sesgado aunque con menor varianza. En estos casos es razonable elegir aquel estimador con menor error promedio de predicción del parámetro (formalmente: con menor error cuadrático medio para el tamaño muestral prefijado, siendo esta cantidad la suma de la varianza del estimador y del cuadrado de su sesgo). Un ejemplo de esto nos lo proporciona la estimación de la varianza σ^2 de una población. En principio, lo más óptimo

siempre existía un estadístico suficiente a la hora de estimar un parámetro, como Fisher se vio obligado a reconocer.

El primer método utilizado para construir estimadores fue el método de los momentos, propuesto por Karl Pearson. La idea era simple: tomar como estimador de la media de la población la media muestral; de la desviación típica de la población, la desviación típica de la muestra, y así sucesivamente. En general,

no es usar la varianza muestral S^2 (que se define como el promedio de las diferencias elevadas al cuadrado de los datos con respecto a la media) sino la «cuasivarianza» o varianza muestral corregida $\hat{S}^2$, que a la hora de promediar, en lugar de dividir por n (el tamaño de la muestra) divide solo por $n-1$. La razón estriba en que al trabajar con muestras se calcula la variabilidad en torno a la media de la propia muestra (no en torno a la media de la población, que es lo que realmente interesa), y ello tiende a subestimar la variabilidad de la población total. Al dividir por $n-1$ se obtiene un valor ligeramente mayor que estima mejor la dispersión de la población, porque el estadístico resultante resulta ser un estimador insesgado. Sin embargo, desde el punto de vista del error cuadrático medio, es mejor emplear la varianza muestral S^2 que la cuasivarianza $\hat{S}^2$. El estimador sin corregir es preferible. Finalmente, cuando se dispone de muestras grandes y no es fácil la obtención de estimadores centrados con alta eficiencia, el requisito mínimo que se exige a un estimador es que sea, de acuerdo con Fisher, consistente, entendiendo por ello que se aproxime, al crecer el tamaño muestral, al verdadero valor del parámetro.



Si equiparamos las estimaciones de varios estadísticos con los disparos de varios tiradores, podemos comprender mejor cuáles son las propiedades que debe cumplir un buen estimador. Los disparos del tirador A no se desvían hacia ninguna dirección en particular, pero se observa que están muy dispersos (lo que representa un estimador insesgado pero no eficiente). Los disparos del tirador B están sesgados hacia la izquierda y, además, dispersos (estimador sesgado y no eficiente). Los disparos del tirador C están poco dispersos pero desviados (estimador sesgado y eficiente). Y los disparos del tirador D están centrados y aglutinados (estimador insesgado y eficiente), lo que constituye la mejor opción.

se igualaban los momentos poblacionales con los momentos muestrales, y se despejaban los parámetros desconocidos. En su artículo, Fisher juzgó que la eficiencia de este método de construcción de estimadores no era la deseada, puesto que muchos no cumplían las propiedades estipuladas. Los estimadores obtenidos por el método de los momentos son consistentes, pero no son, en general, eficientes (centrados con varianza mínima). La

ventaja de estos estimadores es, desde luego, la simplicidad. Su inconveniente es que al no tener en cuenta la distribución de la población que genera la muestra, no utilizan toda la información disponible.

Desde entonces, Fisher siempre se refirió al método de los momentos de Pearson como «ese método tradicional pero ineficiente». En su ceguera, Karl Pearson nunca se dio por vencido e, incluso, en el que sería su último artículo (publicado póstumamente en 1936 en *Biometrika*), defendería a capa y espada las virtudes de su método, comenzando el texto con la siguiente pregunta retórica: «¿Perdiendo el tiempo ajustando curvas mediante el método de los momentos, eh?».

Un procedimiento que proporcionaba estimadores con buenas propiedades, especialmente en muestras grandes, era el método de máxima verosimilitud, que patentó Fisher y que en germen se encuentra en su primer artículo publicado, de 1912. El precedente más directo del método de máxima verosimilitud se halla en Gauss, aunque también en Daniel Bernoulli, pero la inferencia bayesiana que impulsó Laplace ensombreció este y otros trabajos. No obstante, Fisher fue mucho más lejos que estos matemáticos en promocionar su uso como método universal de construcción de estimadores.

Para entender la noción de función de verosimilitud, que Fisher reintrodujo y es una de las más importantes de la inferencia, hay que distinguir con nitidez dos conceptos muy parecidos. Sea θ el parámetro poblacional desconocido y representemos por X la muestra extraída aleatoriamente de la población. Por un lado se tiene la probabilidad de obtener la muestra X condicionada a cierto valor de θ que se supone conocido, lo que se denota como $P(X|\theta)$ (con X variable y θ fijo) y determina la probabilidad de aparición de cada muestra. En cambio, en un problema de estimación, tenemos una cosa muy distinta: se ha observado la muestra X pero θ es desconocido. Sin embargo, la función anterior sigue siendo útil, ya que si sustituimos X por el valor observado, $P(X|\theta)$ proporciona, para cada valor de θ, la probabilidad de obtener el valor muestral X. Cuando variamos θ, manteniendo X fijo, se obtiene una función que se llama *función de verosimilitud* y se designa como $L(\theta|X)$, con X fijo y θ variable. Conviene advertir

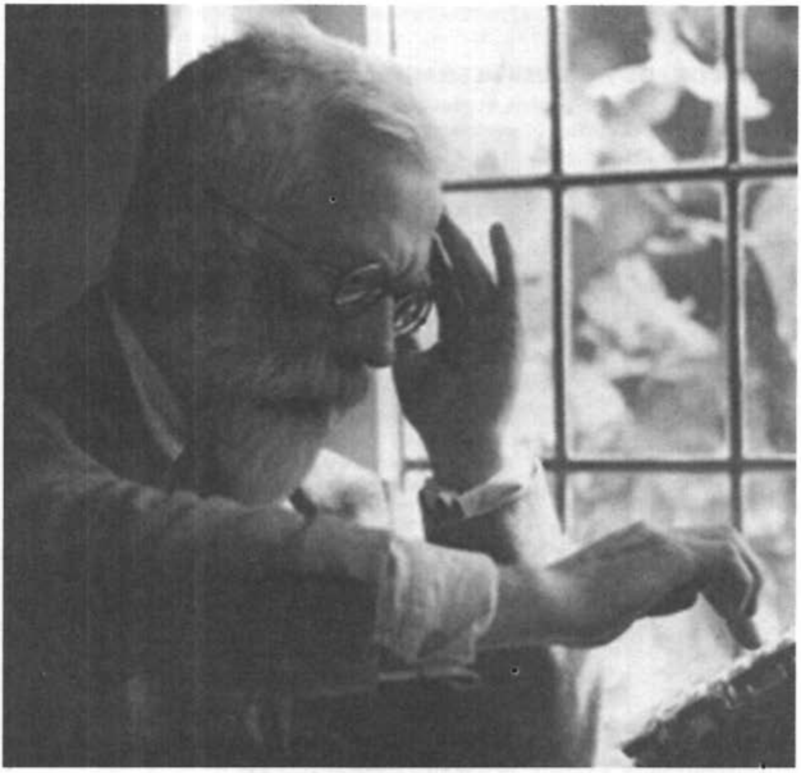




FOTO SUPERIOR:
Una de las fotografías icónicas de Ronald A. Fisher: trabajando con su máquina de calcular, la llamada *Millonaria*. (Fuente: fotografía tomada por Antony Barrington-Brown, reproducida en J.F. Box, *R.A. Fisher: The Life of a Scientist*, Nueva York, Wiley, 1978.)

FOTO INFERIOR:
El Rothamsted Research, antes llamado Estación Agrícola Experimental de Rothamsted, uno de los centros de investigación en agricultura más antiguos del mundo, donde Fisher tuvo ocasión de realizar los experimentos que le permitirían elaborar el corpus de la teoría estadística.

EL PROBLEMA DE LOS TANQUES ALEMANES



Durante la Segunda Guerra Mundial, la producción de Panzers alemanes fue estimada con gran precisión por los estadísticos aliados.

Los estadísticos que durante la Segunda Guerra Mundial trabajaban para los Aliados se toparon con un problema peliagudo: ¿cómo estimar el número total de tanques fabricados por los alemanes a partir de los números de serie de los tanques capturados? Suponiendo que los tanques alemanes habían sido numerados secuencialmente desde 1 hasta N, se trataba de construir un estimador para N. Supongamos, por simplificar, que los tanques capturados tenían los siguientes números de serie: 2, 3, 7, 16. A partir de esta muestra se deseaba estimar N, es decir, el tamaño total de la población de tanques alemanes. Por el método de los momentos, para calcular un estimador de N se igualaba el primer momento poblacional, es decir, la media poblacional:

$$\mu = \frac{N+1}{2},$$

donde se suma 1 porque no hemos empezado a contar desde 0, con el primer momento muestral, es decir, la media muestral, que es:

que, como consecuencia de haber invertido el papel de X y θ de acuerdo al cambio de óptica que se asume en la inferencia, la función de verosimilitud ya no tiene por qué ser una distribución de probabilidad, de modo que —como Fisher no dejó de apuntar— no obedece las reglas del cálculo de probabilidades (una vez se sustituyen los valores particulares de la muestra). Esta función representa el estado de nuestra información con respecto al parámetro de la población. En efecto, en lugar de suponer que conocemos θ y calculamos las probabilidades de observar distintas muestras X, suponemos que hemos observado una muestra X concreta y evaluamos la verosimilitud de los posibles valores de θ. La función de verosimilitud es la herramienta clave para juzgar

$$\bar{X} = \frac{2+3+7+16}{4} = 7.$$

Igualando ambos valores y despejando *N*, se obtenía que la estimación era 13. Sin embargo, por lógica, si en la muestra había salido seleccionado el tanque número 16, era obvio que un mejor estimador era el valor máximo observado en la muestra, 16. Los alemanes habían producido, por lo menos, 16 tanques. No obstante, si solo se consideraba el máximo en la muestra, la estimación tendía a subestimar el tamaño total de la población, puesto que el máximo podía ser igual o menor pero nunca mayor que *N*. En verdad, el mejor estimador posible venía dado por el estimador eficiente (insesgado de mínima varianza) cuya fórmula para *N* era:

$$m + \frac{m-n}{n},$$

donde *m* era el mayor número de serie observado y *n* el tamaño muestral. Esta fórmula puede entenderse como la suma del máximo en la muestra más el «hueco medio» en la muestra. Al valor mayor se le añade el promedio de los huecos entre las observaciones que tenemos, pensando que a continuación suyo debe de haber tantos elementos como más o menos hay entre los valores de que disponemos. En nuestro ejemplo, la mejor estimación para *N* sería:

$$16 + \frac{16-4}{4} = 19 \text{ tanques en total.}$$

la compatibilidad entre los valores muestrales observados y los posibles valores del parámetro.

La intuición de Fisher radicó en escoger como estimación de θ aquel valor que haga máxima la probabilidad de aparición de los valores muestrales efectivamente observados. En otras palabras: se trata de seleccionar como estimador del parámetro aquel valor que maximiza la probabilidad de lo efectivamente ocurrido, de observar los datos que realmente fueron observados. Esto conduce a determinar el máximo de la función de verosimilitud, de manera que se elige como estimador de θ aquel valor que otorgue valor máximo a la función $L(\theta|X)$. Bajo ciertas condiciones de regularidad, los estimadores máximo-verosímiles son asintóticamente

UNA MONEDA TRUCADA

Consideremos una moneda de la que se desconoce la probabilidad p de que al lanzarla salga cara. La moneda se lanza cuatro veces y se obtiene la siguiente serie: *CXCC* (cara-cruz-cara-cara). Por el cálculo de probabilidades sabemos que

$$P(CXCC|p) = p^3(1-p).$$

Por tanto, la función de verosimilitud es:

$$L(p|CXCC) = p^3(1-p).$$

Esta expresión nos permite intuir, por ejemplo, que el valor 0,6 para p es más verosímil que el valor 0,5 dado que $L(0{,}6|CXCC) = 0{,}0864$ y $L(0{,}5|CXCC) = 0{,}0625$. La función de verosimilitud permite discriminar qué valores del parámetro p son más verosímiles a la luz de los datos disponibles. Mediante un cálculo no excesivamente complejo puede demostrarse que la función de verosimilitud alcanza su máximo para el valor 0,75. Nuestra estimación a partir de la muestra observada sería, en consecuencia, que p = 0,75. En esencia, esta es la base del método de estimación de parámetros por máxima verosimilitud.

centrados y eficientes (conforme crece el tamaño de la muestra el sesgo tiende a cero y la varianza a su mínimo) y suficientes (si existe un estadístico así para el problema concreto bajo estudio).

Este método era el que Fisher había empleado en el artículo de 1915 que Karl Pearson había criticado con extrema dureza. Nada tenía que ver con el teorema de Bayes. Para estimar el coeficiente de correlación ρ de toda una población, Fisher había elegido aquel valor que maximizaba la probabilidad de obtener el coeficiente de correlación r observado en la muestra, es decir, el máximo de la función de verosimilitud.

La noción de modelo estadístico, los tres tipos de problemas en inferencia (especificación, estimación, distribución), los tres criterios de estimación (consistencia, eficiencia, suficiencia) y el método de máxima verosimilitud aportaron el marco para el programa de investigación que ha dominado la estadística teórica o matemática durante todo el siglo XX, aunque el carácter vago y

elusivo de muchas de las demostraciones dadas por Fisher dio bastantes quebraderos de cabeza a los matemáticos de las décadas siguientes. La aparición de esta celebrada memoria de Fisher abrió, desde luego, una nueva era en la estadística, consagrando una larga serie de términos (parámetro, estadístico, estimador, etc.) que desde entonces forman parte ineludible de la literatura estadística.

«MÉTODOS ESTADÍSTICOS PARA INVESTIGADORES»

Entre los veranos de 1923 y 1924, Fisher escribió *Métodos estadísticos para investigadores*, un libro que vio la luz en 1925 y hasta la fecha ha sido reeditado en catorce ocasiones. Se trata de su obra más influyente y popular. Da la impresión de ser más un manual para aprendices que un libro de texto, a tenor del estilo persuasivo y la característica ausencia de demostraciones matemáticas. Tal vez en esto radicó su gran éxito. Problemas prácticos, técnicos, teóricos y filosóficos se discuten en el libro a través de ejemplos numéricos muy ilustrativos. Fisher fue un gran matemático aplicado, pero concebía la estadística como una disciplina que no solo necesita del razonamiento deductivo típico de las matemáticas, sino también del razonamiento inductivo que sabe hacer el científico experimentado a partir de los datos que maneja.

El libro contenía una introducción al tema, en la que Fisher mantenía que la estadística no era sino matemática aplicada a los datos observacionales. La estadística se interesaba por el estudio de poblaciones de individuos, moléculas o medidas, fijándose en su variabilidad y en la posibilidad de reducir o simplificar los datos de partida, de extraer toda la información relevante que contuvieran sobre la población subyacente. En su examen de las muestras disponibles, el estadístico realizaba inferencias sobre la población total, pero estas no debían venir expresadas —según subrayaba Fisher con tono agresivo— en el lenguaje de la probabilidad (como querían los partidarios del teorema de Bayes y los métodos inversos de probabilidad) sino, en todo caso, en el lenguaje de la verosimilitud.

A través de los capítulos del libro, Fisher recorría lo que actualmente comprende un curso básico de inferencia estadística. Es de destacar que el autor comenzaba apoyándose en el uso de diagramas. A su entender, su observación no probaba nada, pero frecuentemente sugería cómo comenzar el análisis. Tras repasar las distribuciones de probabilidad fundamentales (normal, binomial y Poisson), presentaba la receta estadística que era la piedra angular de la obra: los «test de significación».

Cada sección del libro dedicada a los test de significación en sus diferentes modalidades (de ajuste, homogeneidad e independencia, para la media, la diferencia de medias o los coeficientes de regresión y correlación) arrancaba con un conjunto de datos con los cuales se había topado en el curso de alguna investigación. Por medio de su disección y explicación, Fisher conducía al lector a través de las diferentes etapas del razonamiento estadístico que llevaban a la solución del problema. El planteamiento de los test estaba basado en el conocimiento de las distribuciones muestrales de poblaciones normales, deducidas con anterioridad por él mismo y otros especialistas en artículos matemáticos que no habían llegado al público de investigadores biológicos o agrónomos. En el libro, Fisher usaba con asiduidad la χ^2 de Pearson, la t de Student y una distribución nueva, que a partir de 1934 sería conocida como la *F de Fisher-Snedecor*, por el matemático estadounidense George Snedecor (1881-1974), que precisó la aproximación logarítmica («log-normal») que en principio empleara Fisher.

Pero, ¿en qué consistía un test de significación? Una prueba de significación constaba, en primer lugar, de una hipótesis nula H_0 que establecía, por ejemplo, que el verdadero valor del parámetro desconocido era tal o cual: $\theta = \theta_0$. La hipótesis de partida del investigador fue bautizada con este nombre por Fisher en 1935 porque en agricultura representaba que no había cambio alguno con el uso de un nuevo fertilizante, que este no tenía efecto, esto es, que la diferencia entre los promedios de crecimiento usándolo y no usándolo era nula.

A continuación, tras delimitar la hipótesis nula que se deseaba poner a prueba, se elegía el estadístico T del test y se calculaba su valor sobre los datos de la muestra X observada, lo que

se denotaba como $T(X)$. Dado que la distribución en el muestreo del estadístico T era conocida, se determinaba la probabilidad de que el estadístico T tomase un valor igual o más extremo que el valor observado $T(X)$ bajo el supuesto de que la hipótesis nula era cierta (es decir, bajo la suposición de que el valor real del parámetro θ era θ_0). Simbólicamente: $P(T \geq T(X)|H_0)$. Este número se denominó *p-valor*. Entonces, si el p-valor era excesivamente pequeño —en general, por debajo de 0,05—, el test se decía que era significativo, porque permitía rechazar la hipótesis nula H_0. En otro caso, el test no era significativo y, para el nivel de significación prefijado de $\alpha = 0,05$, no podía rechazarse la hipótesis nula H_0 y se aceptaba provisionalmente.

«Todo experimento se plantea a fin de dar a los hechos una posibilidad de refutar la hipótesis nula.»

— FISHER, *EL DISEÑO DE EXPERIMENTOS* (1935).

La hipótesis nula solo se rechazaba si la probabilidad de observar una muestra como la dada era demasiado baja. El razonamiento estadístico se basaba en la siguiente disyunción lógica: «o bien ha ocurrido un suceso excepcional (muy improbable), o bien la hipótesis nula no es correcta», empleando palabras del propio Fisher. El p-valor o probabilidad de significación —que en la época no siempre era fácilmente computable— funcionaba para Fisher como una suerte de medida de la evidencia en contra de la hipótesis nula: cuanto menor fuese, más evidencia en contra de la hipótesis se disponía. Un valor demasiado pequeño indicaba que la muestra observada se separaba de lo esperado mucho más de lo que sería achacable al azar, a las circunstancias del muestreo aleatorio, y por tanto el investigador se encontraba ante una hipótesis nula inverosímil, descartable.

Pongamos una ilustración sencilla para fijar ideas. Supongamos que suministramos un nuevo fertilizante a 20 plantas y observamos su crecimiento durante cierto período de tiempo, de manera que medimos si con el nuevo fertilizante han experimentado un aumento (+) o una disminución (–) en el ritmo de crecimiento

con respecto al que tenían antes de usarlo. Nuestra hipótesis nula es que el fertilizante no tiene efecto positivo alguno, es decir, que la distribución entre los aumentos (+) y las disminuciones (–) va a ser completamente azarosa, como si se tratara de las caras y las cruces obtenidas al lanzar una moneda legal, perfectamente simétrica. Por consiguiente, de acuerdo con la hipótesis nula H_0, la probabilidad de + será igual a la probabilidad de –, esto es, $\theta = 0{,}5$. Imaginemos que, tras realizar el experimento, observamos 16 + y solo 4 –. Si elegimos como estadístico T del test el número de + obtenidos, resulta que la probabilidad de obtener 16 + o más bajo el supuesto de que la probabilidad de observar un aumento es de 0,5 es, según puede calcularse fácilmente (véase la tabla siguiente), de solo 0,006. Formalmente: $P(T \geq 16|H_0) = 0{,}006$. Como este p-valor es inferior al umbral de $\alpha = 0{,}05$, el test es significativo y podemos rechazar la hipótesis nula de partida: hay evidencia empírica en contra de la hipótesis de que el fertilizante no tenía efecto, es más, todo parece apuntar a que estimula el crecimiento de las plantas.

Número de +	Probabilidad
16	0,004621
17	0,001087
18	0,000181
19	0,000019
20	0,000001
Suma	**0,006**

Tabla que resume el cálculo de las probabilidades de obtener de 16 a 20 + de acuerdo a la fórmula de la probabilidad binomial:

$$P(\text{el número de + sea } k) = \binom{20}{k} \cdot 0{,}5^{20}.$$

Frente a la creencia común en su entorno, Fisher apuntaba que era el p-valor y no el valor concreto $T(X)$ del estadístico del test lo que constituía una medida del sustrato racional en contra de la hipótesis nula. Así, por ejemplo, el valor particular del estadístico χ^2 calculado para medir la discrepancia entre una serie de valores teóricos y los datos observados no permitía cuantificar

el grado de asociación entre ambas series de valores (lo que sí haría el coeficiente de correlación), porque un mismo valor del estadístico podía ser significativo para una muestra grande pero insignificante para una muestra pequeña. Además, Fisher alertó de que el nivel de significación α no había de ser fijo, rígido. Pero la advertencia pronto cayó en el olvido y se generalizó el uso de 0,05, al punto de no considerar significativo un p-valor de 0,051 y sí otro de 0,049. La elección de este valor frontera no es una cuestión matemática, fijada universalmente, sino que depende del contexto pragmático: si se trata de la prueba de un nuevo fármaco, un nivel de significación del 0,05 implica que se corre un riesgo del 5% de afirmar que el fármaco es eficaz cuando en realidad no lo es (en este caso, como en otros, un nivel del 0,01 o 0,001 puede ser mucho más adecuado).

En suma, los test de significación ideados por Fisher eran, en el fondo, una especie de *modus tollens* estadístico. El *modus tollens* tradicional poseía la siguiente estructura:

Si A, entonces B.
No B.
Luego, no A.

Y la nueva versión estadística era:

Si la hipótesis nula H_0 es correcta, entonces los datos observados no serán estadísticamente significativos al nivel $\alpha = 0,05$ con una alta probabilidad de $1-\alpha=0,95$.

La muestra observada X es estadísticamente significativa al nivel $\alpha=0,05$.

Luego, la hipótesis nula H_0 no es correcta.

Ahora bien, la principal diferencia entre el razonamiento lógico y el razonamiento estadístico es que este último es falible, en el sentido de que no siempre es seguro, pues puede fallar, ya que existe una probabilidad de 0,05 de que por error se rechace la hipótesis nula siendo en verdad correcta. Para sus críticos, esta es la peculiaridad que hace que los test de significación

carezcan de fuerza lógica. Podemos rechazar la hipótesis nula y que, sin embargo, sea verdadera. Los test de significación no podrían, por tanto, inferir la falsedad o la verdad de la hipótesis de partida. Fisher estaría confundiendo los sucesos improbables con sucesos imposibles. No obstante, lo que diferencia a la estadística de la adivinación es, reiterando lo dicho al principio del capítulo, la capacidad para cuantificar con precisión esta probabilidad de error.

Fisher describía los test de significación como un procedimiento para rechazar la hipótesis nula, que en ningún caso podía ser probada o establecida definitivamente. Este planteamiento refutacionista era coherente con la corriente falsacionista que poco después encabezó el filósofo de la ciencia Karl Popper (1902-1994). Tanto para el estadístico como para el filósofo, la ciencia se caracterizaba por el planteamiento de pruebas empíricas que pudiesen refutar o falsar las teorías que conjeturan los científicos. No deja de ser sorprendente que el libro *El diseño de experimentos* de Fisher, que ahonda en este tema y del que hablaremos más abajo, se publicara el mismo año, 1935, en que Popper dio a la imprenta su obra maestra: *La lógica del descubrimiento científico* (aunque el filósofo nunca citó al estadístico). La propuesta metodológica de Fisher era una especie de falsacionismo aplicado a la estadística: se trata de rechazar aquellas hipótesis para las cuales las observaciones sean relativamente inverosímiles (aunque la decisión de rechazar es, desde luego, revisable sobre la base de nuevos hechos). Aquello que distanciaba al estadístico británico del filósofo vienés era que, para nuestro protagonista, los test de significación, aunque metodológicamente deductivos (si tal, tal; no tal, *ergo* rechazamos H_0), eran inductivos por su contenido, pues permitían aprender de la experiencia, aunque siempre de una manera provisional. La hipótesis nula nunca se confirmaba, pero era posible refutarla. Si el test era significativo, la hipótesis era implausible a la luz de los datos; y si no lo era, no indicaba más que la hipótesis era compatible con los datos. No rechazar no quería decir, salvo que se tratara de una batería de test sucesivamente no significativos, aceptar. Ningún experimento aislado demostraba para Fisher una ley natural.

Como ampliaremos en el capítulo 5, la aproximación fisheriana presentaba algunas lagunas. En muchas ocasiones, la evidencia en contra de la hipótesis nula sugería evidencia a favor de cierta hipótesis alternativa, que Fisher no tomaba nunca en consideración dentro de los test de significación. Asimismo, el matemático inglés no hacía demasiado hincapié en el cálculo y la importancia de las probabilidades de error. Finalmente, otra dificultad que salía al paso era la cuestión técnica de qué estadístico elegir para cada test. Una elección, ciertamente, subjetiva, aunque bastante estandarizada. Fisher afirmó que había que agarrarse al principio de suficiencia, eligiendo un estadístico suficiente, es decir, como vimos, un estadístico que contuviera toda la información relevante de la muestra. Pero, desafortunadamente, la mayoría de estadísticos que Fisher empleaba en su libro no cumplían esa propiedad tan deseable (como, por ejemplo, el estadístico χ^2).

A mediados de 1929, Egon S. Pearson (1895-1980), hijo de Karl Pearson y prometedor estadístico por aquel entonces, publicó una reseña sin firmar de la segunda edición del libro en *Nature* que puso furioso a Fisher. Las relaciones entre Pearson hijo y Fisher no volvieron a ser cordiales. Probablemente, este último pensó que Pearson padre estaba malmetiendo detrás. La principal crítica formulada por Egon era que Fisher siempre presuponía que la población subyacente era normal, y la exactitud de los test se venía abajo si esa premisa no era cierta. Curiosamente, Student le había insistido a Fisher sobre este tema por carta, pero este le había hecho oídos sordos. Sería Egon Pearson, espoleado también por Student, el que mediante simulación, es decir, mediante tablas de números aleatorios, probara que muchos test basados en el conocimiento de las distribuciones en el muestreo de poblaciones normales podían seguir empleándose, porque la omnipresente distribución t era robusta, estable aun si desaparecía el supuesto de normalidad. Una actuación emparentada con la que en su día Student usara para comprobar la adecuación empírica de su distribución t, aunque este último no disponía de tablas al efecto y hubo de conformarse con barajar cartas con números extraídos de la medida de la estatura y la longitud del dedo corazón

«ANOVA»

Aparte de las pruebas de significación, el libro de Fisher presentaba el análisis de la varianza, otra novedosa técnica estadística, conocida mundialmente por sus siglas en ingles: ANOVA. Mediante los test de significación se podía comparar la efectividad de un fertilizante con respecto a no usarlo o a otro distinto. Es lo que en la jerga estadística se conoce como *test sobre la diferencia de medias* (en el anexo al final del libro se presenta un ejemplo numérico). Pero, ¿cómo proceder si queremos comparar tres o más fertilizantes, es decir, poner a prueba la hipótesis de que tres o más medias son iguales? Una primera respuesta, bastante ineficiente, sería comparar los efectos de los tres fertilizantes A, B, C dos a dos: A y B; A y C; B y C. Pero, para un nivel de significación fijo de $\alpha=0{,}05$, hacer tres pruebas incrementa la probabilidad de error más allá de lo tolerable: P (algún error en los tres test) $=1-P$ (ningún error en los tres test) $=1-0{,}95^3=1-0{,}86=0{,}14$. La probabilidad de cometer algún error a la hora de rechazar la hipótesis nula de que no hay diferencias es de casi tres veces lo esperado: de 0,14 en vez de 0,05. Si en lugar de tres fertilizantes fuesen cuatro, habría que realizar seis pruebas, lo que empieza a ser demasiado costoso. Para solventar estos escollos, Fisher ideó el análisis de la varianza, que mediante la comparación de las varianzas muestrales —de ahí el nombre— permite sacar alguna conclusión sobre los valores relativos de las medias poblacionales. Supongamos que se han rociado seis parcelas con tres tratamientos diferentes A, B y C (dos parcelas para cada fertilizante). Se observa el rendimiento de cada parcela y se calcula el promedio de productividad de cada tratamiento:

Tratamiento A	Tratamiento B	Tratamiento C
11	6	1
9	5	3
Media A = 10	Media B = 5,5	Media C = 2

A continuación, se calcula la gran media, la media total:

$$\text{Media} = \frac{11+6+1+9+5+3}{6} = 5{,}83.$$

En el experimento se pueden identificar tres tipos de variabilidad: la variación total entre las 6 parcelas (cada una tuvo rendimientos diferentes); la variación entre tratamientos (A, B y C no tuvieron el mismo rendimiento), y la variación dentro de cada tratamiento debida al error o al azar, también llamada *variabilidad interna* o *residual* (no todas las parcelas tratadas con A tuvieron el mismo resultado). La comparación entre estas fuentes de variación permite discriminar la igualdad de efectos de A, B y C. Si la variación entre tratamientos no es del mismo orden que la variación dentro de cada tratamiento, es razonable suponer que la diferencia

sea achacable a los distintos efectos de A, B y C. Y si esta diferencia es estadísticamente significativa, la hipótesis nula podrá rechazarse. Esta diferencia entre la variación «entre» tratamientos y la variación «dentro» de cada tratamiento es, precisamente, lo que mide el análisis de la varianza mediante el cociente de varianzas, cuya distribución es la F de Fisher-Snedecor. Parte del éxito del análisis de la varianza se debe a su presentación en forma de tabla. Como la variación total es igual a la suma de la variación de cada tratamiento más la variación debida al error dentro de cada tratamiento, la suma de cuadrados total (SCT) —de cada observación respecto de la gran media— puede descomponerse como la suma de cuadrados de los tratamientos (SCTR) —entre la media de cada tratamiento y la gran media— más la suma de cuadrados del error (SCE) —de cada observación respecto de la media de su tratamiento—:

$$SCT = (11 - 5{,}83)^2 + (6 - 5{,}83)^2 + \dots + (3 - 5{,}83)^2 = 68{,}83.$$
$$SCTR = 2\cdot(10 - 5{,}83)^2 + 2\cdot(5{,}5 - 5{,}83)^2 + 2\cdot(2 - 5{,}83) = 64{,}33.$$
$$SCE = (11 - 10)^2 + (6 - 5{,}5)^2 + \dots + (3 - 2)^2 = 4{,}5.$$
$$SCT = 68{,}83 = 64{,}33 + 4{,}5 = SCTR + SCE.$$

Después de obtener las sumas de cuadrados, hay que calcular los promedios respectivos, dividiendo cada cantidad por su número de grados de libertad, es decir, por el número de datos menos 1. En nuestro ejemplo, *SCT* se divide entre 6-1=5 (ya que había 6 observaciones); *SCTR* entre 3-1=2 (ya que eran 3 tratamientos), y, finalmente, *SCE* por el número de grados de libertad que salen de despejar en la igualdad *SCT=SCTR+SCE*. Esto es: *SCE=SCT-SCTR*=5-2=3, que coincide con la diferencia entre el número de observaciones y el número de tratamientos. El cálculo de los cuadrados medios lo resumía Fisher en una tabla como la siguiente, donde también se calculaba el valor de la razón *F* entre los cuadrados medios de los tratamientos y del error:

Tabla ANOVA				
Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***F***
Entre tratamientos	64,33	2	32,17	21,44
Dentro de tratamientos	4,50	3	1,5	
Variación total	68,83	5		

Por último, como el p-valor o probabilidad de que una distribución *F* con 2 y 3 grados de libertad tome un valor igual o superior a 21,44 es, según se muestra en las tablas, de 0,02, que es menor que 0,05, puede rechazarse la hipótesis de que los tres fertilizantes actúan de igual manera. Es más, según los datos parece que el fertilizante A es, pese al poco tamaño de la muestra, el más beneficioso.

de 3000 criminales. Actualmente, una variante de este método recibe el luminoso nombre de *método de Monte-Carlo*.

La circulación de *Métodos estadísticos para investigadores* dictaminó el fin de la edad de la correlación y el ajuste de curvas. Hasta Fisher, los estadísticos dedicaban la mayor parte de sus esfuerzos al cálculo de coeficientes, siguiendo el ejemplo de Karl Pearson. Pero una confusión crucial permeaba toda su investigación. En general, no distinguían entre el problema de la estimación del valor del coeficiente, es decir, del grado de asociación entre dos o más variables, y el problema adjunto de testar la significación de esta asociación, su posible existencia. Además, Fisher revitalizó, frente a la escuela abanderada por Pearson, el empleo de muestras de tamaño modesto, transformando los métodos estadísticos en algo vivo, rotundo y bien trabado.

«EL DISEÑO DE EXPERIMENTOS»

En la última sección de *Métodos estadísticos para investigadores*, Fisher discutía y ejemplificaba el diseño de experimentos en agricultura, un campo a medio camino entre el laboratorio y el invernadero con el que se había familiarizado gracias a su estancia en Rothamsted. Poco después, dentro de un artículo publicado en 1926, perfilaba aún más las líneas maestras que debían regir cualquier experimento. La tormenta de ideas precipitó en otro *best-seller*: *El diseño de experimentos*, que salió de la imprenta en 1935 y en el que Fisher recogió los principios básicos del diseño experimental tal y como los había pergeñado durante los años veinte. Esta obra innovadora conoció ocho reediciones, y se trata más bien de un libro de ideas que de cálculos, que ha tenido una gran repercusión en la investigación agraria y, en general, experimental.

La estadística, según enseñó Fisher, es necesaria para saber cómo implementar pruebas que respondan a preguntas del siguiente cariz: ¿qué fertilizante es mejor?, ¿cuál de estos medicamentos es más eficaz?, etcétera. A veces no es posible contestarlas mediante estudios concretos que analicen la acción del

fertilizante o del medicamento en el metabolismo de la planta o del organismo en cuestión, sino que es más seguro recoger datos y comparar resultados. Ahora bien, la recogida de datos puede llegar a ser un proceso de lo más arduo tanto si el experimento encaminado a producirlos no se ha diseñado con cuidado como si el científico no es ducho en interpretar su resultado. De la primera falla, como aclaraba Fisher, se ocupa el diseño de experimentos. De la segunda, la lógica de la inferencia científica. Para el estadístico, el diseño y la lógica son las dos caras de la misma moneda.

La exploración del mundo biológico requiere obligatoriamente de la realización de experimentos controlados. No basta con la observación pasiva. Las técnicas de muestreo consisten en observar una muestra representativa de la población, anotando los valores de las variables bajo estudio. Por el contrario, el diseño de experimentos fija ciertas variables y observa la respuesta en otras, midiendo los cambios que inducen. Cuando los datos se obtienen mediante un adecuado diseño experimental, se tiene una base empírica más sólida para juzgar las relaciones que median entre las variables.

Los objetos que reciben el «tratamiento» —un nombre, ligado al uso de fertilizantes, que ha perdurado— son las unidades experimentales. En el caso de un experimento agrícola, las unidades experimentales son las parcelas o las variedades de plantas tomadas en consideración. Por su parte, el factor es la variable cuyo impacto en tales unidades desea medirse. Cualquier experimento bien planeado debe fijarse, siguiendo a Fisher, no solo en la comparación entre los distintos tratamientos, sino también en poner a prueba la significación de las diferencias observadas por medio de un test estadístico. En consecuencia, todos los tratamientos han de aparecer al menos por duplicado y, preferiblemente, repetidos varias veces. Si queremos comparar los tratamientos A y B, lo idóneo es aplicarlos simultáneamente sobre varios pares de parcelas. Jugárselo todo a una carta, a un único par de parcelas, es demasiado arriesgado y puede conducir a conclusiones erróneas, ya que la muestra no tiene por qué ser representativa. Pudiera ser que la diferencia observada entre los tratamientos A y B se

debiera simplemente a la distinta fertilidad de la tierra de cada parcela y no, pongamos por caso, a que A fuera más beneficioso que B. El principio de repetición o replicación formulado por Fisher servía, por tanto, para acotar el error experimental, es decir, la variación aleatoria o azarosa que escapa al control del experimentador (como que los suelos de las parcelas sobre las que se ha aplicado A y B tengan distinta fertilidad).

«Consultar a un estadístico después de que haya concluido el experimento es, muy a menudo, pedirle que realice un examen post-mortem. Quizá pueda decir de qué murió el experimento.»

— Intervención de Fisher en el Primer Congreso Indio de Estadística (1938).

En la tesitura de diseñar un experimento el científico ignora un sinfín de factores que pueden influir en el resultado. Es incapaz de dominar todas las causas que pueden estar operando detrás. Así, por ejemplo, si desea probar un nuevo fertilizante, no es sensato comparar el crecimiento de las plantas a las que se le va a suministrar en un invernadero con el de plantas de años anteriores o de otros invernaderos, que han podido crecer o están creciendo en ambientes desiguales. Lo suyo es comparar el crecimiento en el mismo invernadero entre dos grupos de plantas: un grupo A al que se le suministra el compuesto químico y otro grupo B —denominado *grupo control*— al que no se le suministra. El científico podría inicialmente inclinarse por plantar los dos grupos de plantas en dos surcos paralelos: el A a la derecha, el B a la izquierda. Pero al hacerlo de este modo podría ser que diversos factores desconocidos —la incidencia solar en cada surco o las corrientes de aire en el interior del invernadero— influyeran en el crecimiento de las plantas enmascarando el verdadero efecto del fertilizante. El instrumento más general para evitar estas desviaciones es lo que Fisher denominó *principio de aleatorización*. Cada pareja de plantas de tipo A y B ha de irse colocando en los surcos de manera aleatoria. Se puede tirar una moneda, de forma que si sale cara, se coloca la primera planta A a la derecha y la primera planta B

a la izquierda. Recíprocamente, si sale cruz, se coloca la planta A a la izquierda y la planta B a la derecha. Y así sucesivamente. Mediante este procedimiento, cualquier diferencia significativa en el crecimiento entre los dos grupos de plantas podrá ser achacada al nuevo fertilizante.

Hasta Fisher, la asignación de tratamientos se realizaba sistemáticamente, lo que podía viciar los resultados. Aleatorizar no cuesta nada y protege contra la influencia de posibles factores conocidos e incluso desconocidos, eliminando las causas de variación fortuita que pueden oscurecer o empañar la evidencia. Sin aleatorizar hubiera podido darse el caso de que el surco seleccionado para plantar el grupo A fuese, sin saberlo, de mayor productividad que el elegido para plantar el grupo B, de manera que la heterogeneidad del suelo camuflase el verdadero efecto del nuevo fertilizante. De hecho, tal y como se habían tomado los datos en Rothamsted, la influencia de las lluvias y de la meteorología en general enmascaraba la posible influencia de los abonos y fertilizantes que se estaban probando en las cosechas. Ambos factores estaban confundidos. Fisher no solo dijo qué andaba mal, sino que explicó cómo hacerlo bien. Inesperadamente, con motivo de la aleatorización como forma de neutralizar factores externos, Fisher estuvo a punto de romper con su viejo amigo Student (aunque el obituario que le escribiría en 1937 se desarrollaría en términos muy elogiosos). Este principio desencadenó bastante controversia, puesto que muchos científicos pensaban que, dado que conocían el material que tenían entre manos, era preferible un experimento sistemático, sin darse cuenta de que con ello condenaban el uso de los test de significación, que requieren de muestras aleatorias.

En ocasiones el diseño completamente aleatorizado de experimentos tropieza con un escollo difícil de salvar: la heterogeneidad de las unidades experimentales (por ejemplo, del terreno de las parcelas). La asignación aleatoria de los tratamientos a las unidades experimentales presupone que todas son homogéneas entre sí. Si esta última condición no se cumple, hay que clasificarlas por bloques (dentro de los cuales se aplicarán aleatoriamente todos los tratamientos, claro). La razón de agrupar en bloques es

evidente: cuanto más heterogéneas son las unidades, mayor es el error experimental y menor la oportunidad de detectar diferencias significativas atribuibles a los diferentes tratamientos. El agrupamiento «bloquea» ese factor externo que provoca una variación en la respuesta que no es de interés, porque no depende de la reacción a los fertilizantes sino, por ejemplo, de las distintas variedades de suelos a los que se les han suministrado. Es lo que Fisher denominó *diseño aleatorizado por bloques*.

Imaginemos que se desea probar cinco tratamientos (A, B, C, D y E) sobre 20 parcelas. Una preparación aleatorizada sería, por ejemplo: B, C, A, C, E, E, E, A, D, A, B, C, B, D, D, B, A, D, C, E, donde cada tratamiento es probado cuatro veces. No obstante, es posible establecer restricciones sobre el diseño completamente aleatorizado del experimento que eliminen parte del efecto debido a la heterogeneidad de la tierra —al «gradiente de fertilidad», como decía Fisher— y, por tanto, incrementen la sensibilidad para detectar diferencias entre tratamientos. Una idea es, prosiguiendo con el ejemplo, dividir las 20 parcelas en 4 bloques según su composición, de manera que en cada bloque aparezca cada tratamiento una vez: AECBD, CBEDA, ADEBC, CEBAD. (Es conveniente respetar la aleatorización dentro de cada bloque para evitar sorpresas.) Así, se reduce la variabilidad final del experimento de manera que es posible estimar la parte que corresponde a las diferencias entre tratamientos con más agudeza.

Tanto en el diseño completamente aleatorizado como en el diseño por bloques, la técnica estadística que proporciona el examen de los datos no es otra que el *análisis de la varianza* o una adaptación suya (ANOVA a una o dos vías). Esta poderosa herramienta creada por Fisher suplía las carencias de algunos de los laboriosos y a menudo erróneos métodos que estaban en boga, y permitía comparar de una vez la acción de más de dos tratamientos —por ejemplo: fosfato, sulfato, clorato o nada— separando las diversas fuentes de variación hasta aislar la del factor que interesaba medir: la debida a la acción de los tratamientos sobre las parcelas.

En resumidas cuentas, Fisher enseñó que los diseños sistemáticos no debían utilizarse. Con un diseño completamente alea-

EL ANTECEDENTE DE LOS SUDOKU

Cuando se desea bloquear el efecto de más de un factor externo que puede provocar resultados equívocos, se emplea el diseño en *cuadrado latino*. Si queremos estudiar el efecto de cinco fertilizantes (A, B, C, D y E), pero se considera que dicho efecto puede estar mediatizado por los tipos de suelo y de insecticidas empleados (supongamos que hay otros cinco tipos de cada uno), un experimento por bloques necesitaría de $5 \cdot 5 \cdot 5 = 125$ unidades experimentales. Obviamente, razones de índole económica desaconsejan experimentar con tantas parcelas. Ante esta situación es posible recurrir a una clase especial de diseño en bloques incompletos aleatorizados: el modelo en cuadrado latino. Este esquema experimental consiste en asignar uno de los factores externos a las filas y el otro a las columnas, de manera que cada tratamiento ocurra una vez en cada fila y en cada columna. Por consiguiente, el número de filas y de columnas ha de ser el mismo: el número de tratamientos. Estamos ante un cuadrado, que se llama latino porque el matemático Leonhard Euler empleó letras latinas para rellenarlo. El popular rompecabezas sudoku no es sino un caso especial de cuadrado latino, en el que no se usan letras sino dígitos del 1 al 9. Este refinado diseño permite al investigador obtener mucha información con una muestra pequeña, ya que elimina la variación extraña mediante el bloqueo simultáneo en los dos factores externos, de manera que las posibilidades de detectar diferencias significativas entre los tratamientos se doblan. En nuestro ejemplo, los 5 tratamientos consabidos se probarían sobre solo 25 parcelas, distribuidas como en el siguiente cuadrado latino:

D	E	C	B	A
B	D	E	A	C
C	A	B	D	E
E	B	A	C	D
A	C	D	E	B

Curiosamente, entre los 56 cuadrados latinos posibles de tamaño 5×5, el llamado *cuadrado de Knut Vik*, basado en el movimiento del caballo de ajedrez, demostró ser más preciso en la estimación que la media del resto de cuadrados latinos. Análogamente, los cuadrados latinos diagonales, aquellos que en la diagonal portan siempre el mismo tratamiento, mostraron ser menos precisos, lo que Fisher interpretó como un argumento más a favor del principio de aleatorización.

torizado, se evitaban los sesgos debidos a la distinta fertilidad de las parcelas, pero el error experimental total podía ser innecesariamente grande. En un experimento bien planeado, ciertas restricciones podían ser impuestas sobre la aleatorización, de manera que la variabilidad debida a la distinta fertilidad de los suelos se eliminara notablemente y fuese más fácil estimar la parte que correspondía a la diferencia entre los tratamientos. Por medio del diseño en bloques, el valor del experimento se incrementaba varias veces, de forma que solo la repetición sucesiva del experimento originario podía igualar la precisión lograda (y esto suponiendo que la replicación fuese factible, ya que en agricultura difícilmente se cuenta con las mismas condiciones meteorológicas).

Otro de los avances que lleva la firma de Fisher es la posibilidad de testar más de un factor de interés en un único experimento gracias a un uso cuidadoso de la estadística, lo que redujo los experimentos diseñados para contrastar un solo factor al plano de los procedimientos ineficientes y costosos. En muchas situaciones prácticas resulta necesario evaluar a un mismo tiempo los efectos de varios factores, así como su posible interacción. Un experimento factorial posee la ventaja de estudiar de golpe dos o más factores en lugar de tener que realizar dos o más experimentos independientes. Más aún, la utilización del diseño factorial identifica la interacción que pueda existir entre los factores, lo que es imposible de detectar si los experimentos se realizan por separado. En el caso de dos factores en que uno tiene tres niveles y el otro dos (por ejemplo, tres niveles de abono con nitrógeno, correspondientes a las dosis factibles, de 0 a 2, y dos niveles de potasio, 0 y 1), tendríamos un experimento factorial con un total de $3 \times 2 = 6$ tratamientos. La respuesta sería observada bajo seis tratamientos diferentes.

Fisher luchó denodadamente contra la máxima, hasta entonces respetada, de variar un único factor en cada ocasión. Hasta que arrumbó esta creencia, la mayoría de investigadores pensaba que lo mejor era investigar un factor cada vez. Sin embargo, la naturaleza, por así decirlo, respondía mejor a un cuestionario bien planeado que a una pregunta aislada.

ZEA MAYS Y LA INFERENCIA ESTADÍSTICA NO PARAMÉTRICA

El tercer capítulo de *El diseño de experimentos* está dedicado al análisis de un célebre experimento llevado a cabo por Charles Darwin con el fin de probar que las plantas obtenidas por fecundación cruzada crecían más que las autofecundadas. Con la ayuda de Galton, Darwin comparaba el crecimiento de 15 pares de plantas de la especie *Zea mays*, es decir, de maíz. El primer miembro de cada par provenía de una fecundación cruzada, mientras que el segundo lo hacía de una autofecundación. Los pares eran plantados simultáneamente en una misma maceta, buscando que las condiciones ambientales —agua, luz, temperatura, etc.— fuesen idénticas para cada uno de los dos. Estas precauciones tomadas por Darwin servían para que se tuviera lo que se denomina una *muestra pareada*, lo que, frente a la posibilidad de tener dos muestras independientes de 15 plantas cada una por su lado, incrementa la sensibilidad del experimento, esto es, su capacidad para detectar diferencias significativas, porque reduce el error experimental. Mediante el test de la *t* de Student (un ejemplo del cual se presenta en el anexo para no entorpecer la lectura), Fisher estudiaba la diferencia en los promedios de crecimiento y concluía que Darwin estaba en lo cierto, aunque no dejaba pasar la ocasión de reconvenirle que no aleatorizara la plantación de cada tipo de planta en una mitad de la maceta. Asimismo, amonestaba a Galton por manipular falazmente los datos de la muestra, reordenándolos a su antojo.

Inferencia no-paramétrica

A continuación, anticipándose a la crítica que ciertos estadísticos teóricos alejados de la práctica experimental (una alusión obvia a Egon Pearson) podían hacer señalando que el uso del test de significación presuponía que los dos grupos de datos eran muestras provenientes de poblaciones normales, Fisher ideaba un método nuevo que conducía a la misma conclusión. Era un ejemplo temprano de lo que sería la inferencia no-paramétrica, una brecha abierta en la inferencia estadística que sería muy explotada tras la Segunda Guerra Mundial, y que se diferencia de la inferencia paramétrica organizada por Fisher en que no especifica nada sobre la forma de la distribución de la población subyacente y los parámetros de que depende. Los test no paramétricos presentan una menor sensibilidad que los test paramétricos, pero no parten de la hipótesis de normalidad, lo que los hace más generales.

Además, *El diseño de experimentos* convirtió el tomar el té en una cuestión estadística. Fisher tenía la costumbre, desde los tiempos de Rothamsted, de tomarlo con todos los miembros de su departamento. Un día, al dar la taza a la doctora Muriel Bristol, esta declinó diciendo que prefería que la leche se vertiera primero. A su

juicio, el té tenía un sabor diferente si la leche se ponía antes o después. Fisher contestó que aquello era irrelevante. William Roach, otro miembro del departamento, quien después se casó con ella, propuso realizar un experimento: irle ofreciendo una serie de tazas mezcladas de diferente manera y comprobar si era capaz de distinguirlas. La doctora identificó todas y cada una de las tazas correctamente. Y Fisher incluyó la historia en su libro como hilo conductor para plantear una serie de interrogantes que sirvieran de guía de acción para enfrentarse a cualquier experimento: ¿cuántas tazas debían servirse?, ¿en qué orden?, ¿cuántas se tenían que acertar?...

Si se le daba una sola taza de cada tipo, la probabilidad de que la doctora acertara al azar era de 1/2, es decir, demasiado alta para discriminar si acertaba por casualidad o porque podía distinguir una mezcla de la otra. Si solo se estaba dispuesto a creerla cuando la probabilidad de que superara correctamente la prueba por casualidad fuese suficientemente pequeña (menor de 0,05, para que este contratiempo ocurriera menos del 5% de las veces), no servía darle 2 tazas de cada tipo, ya que por casualidad acertaría 1 de cada 6 veces (hay 6 formas de elegir 2 entre 4 objetos y solo una es la correcta), es decir, el 17% de las veces. Tampoco funcionaba ofrecerle 3 tazas de cada tipo, ya que acertaría por casualidad 1 de cada 20 veces (hay 20 formas de seleccionar 3 objetos entre 6). Lo que arrojaba una probabilidad que es igual pero no inferior al límite estipulado de 0,05. En cambio, si se le daban 4 tazas de cada tipo, la probabilidad de acertar por azar era solo de 1 entre 70 (existen 70 maneras distintas de elegir 4 objetos entre 8), es decir, de 0,014, de modo que si la doctora acertaba en estas condiciones se podía afirmar que sí sabía distinguir una preparación de otra. Esa era la raya que al trazarla permitía distinguir si solo adivinaba el resultado o verdaderamente estaba capacitada para discernir cómo se había preparado el té.

Adicionalmente, Fisher recalcaba que las tazas debían presentársele a la doctora en un orden aleatorio, para que el experimento estuviera bien diseñado y el test de significación fuese aplicable. Con este maravilloso ejemplo de experimento psicofísico, el estadístico inglés arrancaba un clásico apabullante que dinamitó la tradición experimental heredada.

LA EMERGENCIA DEL RAZONAMIENTO ESTADÍSTICO

Fisher revolucionó la investigación experimental, describiendo métodos, hoy de uso corriente, para exprimir al máximo los experimentos con muestras pequeñas, evitando en lo posible la penetración de factores extraños. Ese niño debilucho con muchas ganas de aprender y dotado de una profunda visión geométrica se convirtió en uno de los científicos que más aportaciones ha hecho a la estadística, sino el que más. En 1929 fue admitido en la Royal Society. Y al retiro de Karl Pearson en 1933, su puesto en el University College de Londres se escindió en dos: una cátedra de Estadística para su hijo Egon y otra de Eugenesia para Fisher, que abandonó Rothamsted para ocupar la «cátedra Galton», aunque Karl Pearson movió todos los hilos para evitarlo. Por descontado, Egon Pearson heredó la antipatía hacia su padre de que Fisher hacía gala, de forma que las hostilidades bajo el techo común no tardaron en desencadenarse, propiciando que la atmósfera entre ambos laboratorios —el biométrico y el eugenésico— fuese irrespirable.

No obstante, para Fisher fueron años placenteros, plagados de éxitos profesionales e intelectuales. Las distinciones acrecentaron su fama, transformándolo en un investigador de prestigio internacional. George Snedecor, con la extraordinaria síntesis que fueron sus *Métodos estadísticos* (1940), así como Harold Hotelling, hicieron mucho por su temprano reconocimiento en América. En Europa, la publicación en colaboración con Frank Yates (1902-1994), su discípulo más aventajado en Rothamsted, de las *Tablas estadísticas para la investigación biológica, agrícola y médica* (1938) contribuyó a difundir sus ideas. No obstante, sería el manual escrito por el matemático sueco Harald Cramér, titulado *Métodos matemáticos de la estadística* (1946), la obra que más ayudaría a expandir su concepción de la estadística, al vincular la inferencia estadística británica con la teoría de la probabilidad continental. En este libro ya aparece, por ejemplo, la cota de Cramér-Rao, deducida tanto por el matemático sueco como por el estadístico indio C.R. Rao (doctorado con Fisher), que acota por abajo la varianza mínima de un estimador, completando la teoría fisheriana.

De resultas de todo ello, se fraguó la definitiva autonomía de los métodos estadísticos, que sedimentaron en torno al concepto de modelo estadístico introducido por Fisher (aunque alguna rama actual de la estadística, como el análisis exploratorio de datos definido por John W. Tuckey en 1977, no lo emplea, razón por la cual a veces se lo considera una *rara avis* dentro de la ciencia estadística). A nuestro juicio, aunque muchos historiadores de la ciencia hablan de la revolución estadística del siglo XIX, creemos que —desde una perspectiva interna— la verdadera revolución se produjo durante los años veinte y treinta del siglo XX, cuando la inferencia estadística sufrió una inyección probabilística y, al tiempo, experimental. Si se drenaran todos los materiales biológicos, sociológicos, etcétera, la estadística —como no dejó de anotar Fisher— se convertiría en una disciplina secundaria. Las aplicaciones son los materiales imprescindibles que hacen de esta ciencia algo más que mera matemática aplicada.

Esta dimensión de la estadística, capaz de proyectar un haz de luz sobre múltiples campos, aceleró su institucionalización —simbolizada con la fundación del Laboratorio Estadístico de Iowa, en Estados Unidos, en 1933 por Snedecor (al que Fisher visitó en varias ocasiones)—, así como su auge durante y después de la Segunda Guerra Mundial, cuando los laboratorios estadísticos se aliaron con las universidades y las industrias en el esfuerzo bélico. Los análisis estadísticos que antes parecían una excentricidad —como los de Galton sobre la eficacia de la oración o la longitud de la soga de la horca— se convirtieron en una realidad cotidiana en econometría, meteorología, epidemiología (la bioestadística), ingeniería industrial (el control de calidad)... Una multiplicación de campos, investigadores, departamentos, libros y revistas especializadas que también se vio empujada por la extensión de los ordenadores, que facilitan el uso de los métodos estadísticos (por ejemplo, para generar números aleatorios sin tener que recurrir a las sempiternas tablas).

En concreto, los test de significación y los principios de experimentación dictados por Fisher han conocido mil y una prácticas exitosas, desde la prueba de fertilizantes a vacunas. Sin ir más lejos, el reciente anuncio de la detección del célebre bosón

de Higgs en julio de 2012 ha tomado el aspecto de un p-valor: los físicos han informado de que la probabilidad de detectar un efecto como el observado en el acelerador de partículas bajo el supuesto de que se trata de mero ruido de fondo (la hipótesis nula) es inferior a 0,0000003, y han interpretado esta significación estadística como una fuerte evidencia para presuponer la existencia de la mencionada partícula (ya que de otra manera no se explica la señal). Un p-valor que todavía se ha hecho más pequeño tras los experimentos reportados en marzo de 2013, dando la razón a las sabias palabras de Fisher en *El diseño de experimentos*:

> Un fenómeno es demostrable experimentalmente cuando se conoce cómo conducir un experimento que raramente falla para darnos un resultado estadísticamente significativo.

Resumiendo: al calor de los experimentos agrícolas, Fisher cerró el grueso de la teoría estadística y, al sembrar la recurrencia de estos métodos, segregándolos de la biometría y otros contextos técnicos, selló la posibilidad de su aplicación continuada y flexible, de manera que la estadística logró irrumpir en todos los órdenes. A la vanguardia de ese ejército de revolucionarios que son los estadísticos siempre figurará Ronald Aylmer Fisher, que puso la piedra mayor del puente que vincula esta disciplina matemática con la práctica experimental.

CAPÍTULO 4

La síntesis entre Darwin y Mendel

Desde los tiempos de estudiante universitario, Fisher se propuso reconciliar a Darwin con Mendel; en otras palabras, la selección natural de las especies con las leyes que rigen la herencia. Sin las aportaciones contenidas en *La teoría genética de la selección natural* (1930), el darwinismo habría permanecido eclipsado y la teoría sintética de la evolución habría tardado años en afianzarse.

Durante su estancia en la Estación Agrícola Experimental de Rothamsted, Fisher no solo tuvo tiempo de refundar la estadística como ciencia matemático-experimental, sino que desarrolló toda una serie de experimentos biológicos encaminados a combinar la teoría de la evolución de Darwin con la teoría de la herencia de Mendel. A pesar de que la estación no estaba oficialmente involucrada en la investigación, le permitió dedicar parte de su esfuerzo a la cría de ratones, caracoles y gallinas, facilitándole tierras para ello (aunque la colonia de ratones era atendida constantemente por su mujer e hijos).

No obstante, su atracción por la materia venía de antes, de mucho antes. Entre 1912 —el año en que publicó su primer artículo— y 1919 —cuando se instaló en Rothamsted—, Fisher escribió casi una centena de textos, de los que más de noventa tenían que ver con temas biológicos y solo el resto con la estadística o las matemáticas. Cabe destacar, entre los dedicados a la biología, su influyente artículo sobre genética de 1918: «La correlación entre parientes bajo el supuesto de herencia mendeliana».

Mientras sufría impartiendo clases a adolecentes, el científico británico comenzó a darle vueltas a una cuestión que había planteado Karl Pearson: ¿era la variación en las poblaciones humanas consistente con el modelo mendeliano de la herencia? En Cambridge, donde los mendelianos predominaban, Fisher se

había convencido de que las leyes de Mendel explicaban la herencia, y quería mitigar el debate entre biómetras y mendelianos mostrando que las mediciones de los primeros eran coherentes con los principios de los segundos. Aunque cada rasgo o factor hereditario —a partir de 1922 Fisher reemplazó el término *factor* por el de *gen*— se ajustaba por separado a las leyes discretas de Mendel, la acumulación de factores hereditarios que se daba en los individuos y en las poblaciones respetaba la ley continua de la selección natural de Darwin, a la manera como la suma de errores en la observación astronómica se distribuye normalmente a pesar de que cada uno de los errores en particular no lo haga así.

Los héroes de juventud de Fisher habían sido Darwin y Ludwig Boltzmann, creador, junto a Maxwell, de la mecánica estadística. En analogía con el conjunto infinito de las moléculas de un gas que estudiaba la mecánica estadística, Fisher imaginaba, tanto en el campo abstracto de la inferencia estadística como en el más práctico de la biología evolutiva, una hipotética población infinita de la que se extraían las muestras observadas. Un artículo posterior de 1922 sobre la dominancia genética especificaba aún más esta analogía pionera:

> La evolución por selección natural puede compararse al tratamiento analítico de la teoría de gases, en el que es posible hacer las más variadas asunciones sobre la naturaleza de las moléculas individuales y, sin embargo, plantear leyes generales sobre el comportamiento de los gases.

El modelo fisheriano de las poblaciones mendelianas era, en suma, una adaptación del modelo de los gases de la mecánica estadística. La variación continua observada en el total de la población podía perfectamente ser el producto de la acción de muchos factores hereditarios discretos.

En el borrador que esbozó hacia 1916, Fisher incorporó por vez primera el término estadístico *varianza*, que definió en la primera página. Asimismo, mencionó de pasada la expresión *análisis de la varianza* como forma de separar la fracción de variabilidad que correspondía a cada causa en la herencia. Pero el núcleo del mismo lo constituía la tesis de que la teoría de Mendel

no se veía rechazada por los datos biométricos. En una carta que le envió a Karl Pearson, fechada en 1916, le decía:

> Recientemente he completado un artículo sobre el mendelismo y la biometría que probablemente sea de tu interés. Me he encontrado con que el análisis de los datos humanos no contradice el mendelismo. Pero el argumento es bastante complejo.

Fisher probó a enviar su artículo a la Royal Society de Londres para que lo publicaran, pero los árbitros expresaron reservas sobre su contenido. Uno de ellos no era otro que Karl Pearson, que aunque no era abiertamente hostil al resultado de la investigación de Fisher, encontró su borrador poco convincente y, probablemente, no entendió del todo las matemáticas empleadas. El otro árbitro fue el biólogo R.C. Punnett, al que paradójicamente Fisher sucedería en el cargo en Cambridge en 1943. Años después, Fisher soltaría el exabrupto de que el artículo había sido referenciado por un estadístico que no sabía biología y por un biólogo que no sabía estadística. En descargo de los árbitros hay que señalar que los artículos de Fisher no siempre eran fáciles de seguir, pues como Student manifestó más de una vez por carta, el *evidently* de Fisher se traducía en varias horas de arduo trabajo para el resto de los mortales.

Finalmente, Fisher retiró el artículo y lo reenvió a la Royal Society de Edimburgo a mediados de 1918, donde fue publicado, no sin dificultad, a su costa, gracias a la ayuda financiera de su amigo Leonard Darwin (1850-1943), hijo de Charles Darwin y quien, desde los tiempos de Cambridge, le apadrinó y sostuvo durante los períodos de penuria económica. El primer paso en pos de la unificación estaba dado.

EL ECLIPSE DEL DARWINISMO

Charles Darwin confirió movimiento a las clases naturales de Linneo. Aunque el dinamismo de Darwin, en contraposición del fijismo de Linneo, flotaba en el aire (ya se encuentra en el trans-

formismo de Lamarck), la originalidad del naturalista inglés reside en haber proporcionado un mecanismo explicativo: la selección natural, entendida como metáfora, según expuso en *El origen de las especies* (1859). El teorema darwiniano de la evolución se basa primariamente en las técnicas de domesticación y cría de animales y plantas (la «selección natural» como extensión de la «selección artificial» practicada por el hombre, pero prescindiendo del sujeto operatorio, del demiurgo selector, y por tanto de cualquier finalidad), y se materializa en los árboles evolutivos que reordenan las especies vivas y los fósiles de las especies extintas (la reconstrucción filogenética de las taxonomías morfológicas).

Durante el período de tiempo que media entre la muerte de Darwin en 1882 y el resurgir de sus ideas en la década de 1930, se produjo un «eclipse del darwinismo» en el que la biología evolutiva se sumió en un estado lamentable de postración, como consecuencia del avance de las teorías mendelianas de la herencia. El trabajo de Mendel fue redescubierto en torno a 1900, treinta y cuatro años después de su publicación y dieciséis después de la muerte de su autor: en el viejo continente, por botánicos como Hugo de Vries, y en las Islas, por William Bateson (a quien se debe la acuñación del término *genética*), que lo empleó como un arma para revalorizar las teorías no darwinianas (lamarckianas o mutacionistas) que defendían una variación no gradual, sino discontinua de las especies. Bateson magnificó las diferencias entre Mendel y Darwin, presentando al primero como hostil a la teoría de la evolución y al segundo como responsable del abandono en que cayó la teoría mendeliana.

La muerte de Weldon en 1906 y de Galton en 1911 dejó prácticamente solo a Karl Pearson en la defensa de la ortodoxia: *Natura non facit saltus*. De hecho, las primeras contribuciones biométricas de Pearson habían consistido en el estudio estadístico de la ley de herencia ancestral de Galton y en la corroboración de la hipótesis de la gradación, mediante la que los biómetras defendían que la evolución no había sido a saltos, como defendían los partidarios de la teoría de la mutación, sino por una selección continua de la variación favorable en la distribución de la descendencia.

DEMASIADO BUENOS PARA SER CIERTOS



Gregor Mendel.

El resultado principal de los experimentos en hibridación de plantas de Mendel fue el descubrimiento de que ciertos caracteres son transmitidos a la descendencia sin atenuación ni fusión, porque son transportados por alguna clase de unidad distintiva o partícula, que Mendel denominó *factores* y nosotros llamamos *genes*. Pero el monje agustino también realizó un contaje exhaustivo de sus experimentos. Así, al cruzar guisantes amarillos con verdes, obtuvo una cosecha en que de 8023 guisantes, 6022 (≈75%) eran amarillos (dominante) y 2001 (≈25%) verdes (recesivo). Se trataba de la segunda ley de Mendel o ley de la segregación. En un artículo publicado en 1936, titulado «¿Ha sido redescubierto el trabajo de Mendel?», Fisher puso de manifiesto, mediante el test de la χ^2, la coincidencia casi total entre los datos observados que publicó Mendel en sus famosos experimentos con guisantes y los resultados teóricos que cabía esperar. Lo más sorprendente es que Mendel había deducido una predicción incorrecta para algunos experimentos y, sin embargo, las observaciones presentaban una similitud notable con esos valores incorrectos. Fisher señalaba que no necesariamente debía haber sido el mismo Mendel quien cocinara los datos, sino algún celoso asistente suyo que no había hecho su trabajo con diligencia y sabía lo que Mendel quería escuchar... El tema, como es natural, levantó gran polémica, y a día de hoy no hay consenso acerca de si Mendel o un ayudante retocaron los datos. A veces poca discrepancia también es sospechosa.

En cuanto bastión de Darwin frente a los embates mendelianos, la escuela biométrica se enzarzó en una dura polémica. En esta oposición férrea influyó, desde luego, la filosofía de la ciencia que asumía Pearson, heredada de sus años de estudiante en Alemania, y que le llevaba a concebir la biometría como mera descripción sin especulación, como una teoría puramente cuantitativa de la evolución natural. Pearson deseaba hacer predic-

ciones probabilísticas sobre la evolución de una línea ancestral, pero sin comprometerse con la discusión metafísica de los mecanismos hereditarios subyacentes. Una meta en consonancia con la biblia del positivismo pearsoniano, *La gramática de la ciencia*, cuyo parecido con la filosofía idealista no dejó de advertir y fustigar Vladimir Ilich Lenin en *Materialismo y empiriocriticismo* (1909). Esta peculiar filosofía fue, por un lado, la que le condujo al desarrollo de una ciencia puramente matemática de la herencia, equipada con herramientas estadísticas para describir los fenómenos observables, pero, por otro lado, la que le obstaculizó valorar la singular aportación presentada por Fisher en 1918. Para Pearson, las poblaciones infinitas y los cúmulos de factores hereditarios de que hablaba Fisher eran inobservables y, por consiguiente, irreales. El disgusto con las imágenes empleadas por Fisher fue mayúsculo.

REVOLUCIÓN EN LA GRANJA

La polémica entre biómetras y genetistas no se cerró, como se ha dicho, hasta que Fisher comprobó que las mediciones empíricas de los organismos concordaban con las leyes postuladas sobre la herencia. El estadístico británico fue el artífice de la síntesis entre Darwin y Mendel, toda vez que demostró que las mediciones eran el resultado de la adición de un gran número de factores mendelianos (los genes) y que los valores experimentales de los coeficientes de correlación se explicaban asimismo por la comunidad de estos factores.

Fisher cumplió con una doble misión. Por un lado, contribuyó significativamente al nacimiento del neodarwinismo, de la teoría sintética de la evolución, en la década de 1930. En esta síntesis confluyeron una multiplicidad de cursos de investigación (biométricos, genéticos, anatómicos, embriológicos, paleontológicos...), como prueba la nómina de autores que participaron en ella: Theodosius Dobzhansky (genetista), Ernst Mayr (zoólogo), George Gaylord Simpson (paleontólogo), etcétera. Por otro lado,

fundó la genética de poblaciones, que es uno de los pilares de la síntesis evolutiva moderna, una disciplina en la que convergen la biología evolutiva y la genética como un todo consistente modelizado matemáticamente.

En este punto, hay que destacar el libro revolucionario que Fisher le dictó a su mujer durante su época en Rothamsted, *La teoría genética de la selección natural* (1930), así como las obras de otros dos grandes genetistas: *Evolución en poblaciones mendelianas* (1931), de Sewall G. Wright (1889-1988), y *Las causas de la evolución* (1932), de J.B.S. Haldane (1892-1964), quien ocupó en 1937 la cátedra de Biometría del University College, asistida con los fondos que la viuda de Raphael Weldon destinó a tal fin al morir. Fisher, Wright y Haldane son los tres tenores de la genética de poblaciones, ya que restablecieron la selección darwiniana como primer mecanismo evolutivo en términos de consecuencia estadística de la genética mendeliana.

«La selección natural no es la evolución.» Con esta categórica afirmación arrancaba el libro de Fisher, que es lo que se llama un clásico de la genética de poblaciones. El aforismo buscaba reclamar la atención sobre el otro componente ineludible de la teoría de la evolución: la genética mendeliana.

Las unidades evolutivas no eran los individuos, sino las poblaciones, cada una con una distribución genética propia. En ausencia de mutaciones, y suponiendo la invariancia del entorno, la evolución de la población más tarde o más temprano cesaría. Pese a que el número de posibles combinaciones de variantes de genes (de «alelos») era inconcebiblemente grande, era finito, de manera que la combinación más adaptada al entorno selectivo terminaría imponiéndose, aunque para ello la selección natural habría de operar sobre las sucesivas generaciones durante un período de tiempo dilatado. Sin embargo, aunque infrecuentes, las mutaciones de hecho ocurrían. Y la historia de la supervivencia del nuevo gen mutante dependía, según ponía de relieve Fisher, tanto de los caprichos de la fortuna como de la ventaja o desventaja selectiva que conllevara en la lucha por la vida.

El razonamiento matemático de Fisher en su libro comenzaba presuponiendo la aparición de un gen mutante en el seno

de una población formada por millones de individuos, y cuya distribución no era otra que la distribución de Poisson o de los «sucesos raros», con media $1+e$ (con $e \geq 0$), donde e representaba la «ventaja selectiva». Si una población presentaba, respecto de un carácter, ejemplares fenotípicamente diferentes (pongamos por caso, polillas blancas y polillas negras), cada uno de los cuales podía corresponder a uno o más genotipos (dependiendo de qué alelo fuera el dominante y cuál el recesivo), de modo que en una generación la proporción observada entre ambos fenotipos era r y en la siguiente, en la descendencia, era $r(1+e)$, entonces e era la ventaja selectiva del alelo que daba lugar a ese fenotipo (por ejemplo, de las polillas negras con respecto a las blancas, que se camuflaban mejor entre el humo de las fábricas inglesas). Naturalmente, la ventaja selectiva e no tiene por qué ser igual a lo largo del tiempo o en distintas condiciones ambientales, de tal forma que lo que es favorable aquí y ahora puede no serlo en otro momento o lugar. En el caso de las polillas, una ventaja selectiva de 0,01 a favor de las polillas mimetizadas con el entorno industrial quería decir que, mientras que la variante blanca dejaba 100 descendientes, la variante negra dejaba 101 (un 1% más).

«En ocasiones he conocido genetistas que me preguntan si es verdad que el gran genetista R.A. Fisher fue también un importante estadístico.»

— Leonard «Jimmie» Savage (1976).

En estas condiciones, Fisher calculó la probabilidad de extinción del mutante en la n-ésima generación. En el caso de no existir ventaja selectiva ($e=0$), la probabilidad de extinción en la sexagésimo tercera generación era igual a 0,9698, es decir, de casi un 97% a favor de la extinción. Sorprendentemente, con una ventaja selectiva del 1% ($e=0{,}01$), la probabilidad señalada era de 0,9591, de casi un 96% a favor de la extinción. Tan solo de un 1% menos. Prosiguiendo con los cálculos, en la 127 generación la probabilidad de no haberse extinguido era de 0,0271 con ventaja

selectiva y de 0,0153 sin ventaja, es decir, el gen mutante tenía casi el doble de probabilidad de supervivencia, aunque ambas probabilidades eran realmente bajas. En el límite, la probabilidad de que la mutación beneficiosa sobreviviera era de cerca del 2% (por su parte, la probabilidad de que lo hiciera la neutra era 0). Ahora bien, si la población era grande, del orden de millones de individuos, habría una cantidad no despreciable de individuos dotados con la mutación benéfica, lo que posibilitaría el cambio adaptativo, sin perjuicio de que muchas mutaciones benignas pudieran perderse por el camino.

Con estos cálculos Fisher también pretendía mostrar cómo la dirección y el sentido de la evolución apenas tenían que ver con los de la mutación, puesto que sin ventaja selectiva el efecto de la mutación en la especie era insignificante y, en el límite, nulo (y esto sin contar con que la mayoría de las mutaciones producen deformidades monstruosas, letales). La selección natural era el proceso por el cual una contingencia improbable como era una mutación veía aumentada gradualmente su probabilidad con el paso del tiempo. La selección natural era, por tanto, el motor principal de la evolución. Lo que le devolvía la razón a Darwin y resucitaba el darwinismo al que tan refractarios habían sido los mendelianos. Las implicaciones biológicas de los resultados matemáticos obtenidos por Fisher fueron extremadamente importantes, y se vieron apoyadas por los experimentos con la mosca del vinagre (*Drosophila melanogaster*, cuyo frenético ritmo reproductor facilita el estudio de mutaciones y cruzamientos).

Además, la obra de Fisher contenía el «teorema fundamental de la selección natural», que santificaba la unión entre Darwin y Mendel, y era la pieza central de la visión de Fisher de la selección natural. Este era su enunciado: «El ritmo de aumento en la adaptación biológica de una población en cualquier momento es igual a la variabilidad genética en adaptación que la población tiene en ese momento». Esta formulación algo críptica hizo de él un elemento oscuro, que tardó bastantes años en ser valorado en su justa medida.

Para que la selección natural pueda actuar sobre un carácter, debe haber algo que seleccionar, es decir, varios alelos, o formas

UN TEMPERAMENTO DIFÍCIL

Ronald Aylmer Fisher estaba dotado de grandes virtudes, pero también poseía notables defectos. Entre ellos, un ánimo belicoso que le llevaba a porfiar y discutir por trivialidades, comportándose en ocasiones con una notoria rudeza tanto oral como escrita dentro de sus controversias con otros colegas estadísticos y genetistas. Ya hemos visto una muestra de ello en su enfrentamiento personal con Karl Pearson, y en el próximo capítulo veremos alguna más a propósito de su concepción de la inferencia estadística o de la relación entre tabaco y cáncer. Esta firmeza en su ideario científico era extensiva a sus creencias religiosas y políticas, teñidas de un claro talante conservador que le llevaba a respetar las tradiciones heredadas de sus padres y denostar cualquier forma de progresismo o comunismo. Fruto de sus convicciones eugenésicas, mantenía que no todos los hombres eran iguales. A todo esto unía algunas de las excentricidades típicas de los matemáticos geniales. Su tendencia a perder papeles importantes o a ser un administrador impaciente y despistado. Por otra parte, su malhadada vista no era óbice para una condición física envidiable, conseguida gracias a que iba corriendo a trabajar a diario. Curiosamente, para poder continuar trabajando en casa con tranquilidad (lo que tenía que ser difícil dada la amplitud de su prole, que constituyó para él una fuente de desesperos monetarios), exigía que siempre hubiera dos puertas cerradas entre él y los niños a fin de poder concentrarse.

Un profesor pésimo

Según todos los testimonios, Fisher fue, sin lugar a dudas, un profesor pésimo, tendente a omitir explicaciones tanto en la docencia como en la investigación. Al respecto, recogemos una anécdota relatada por el estadístico escocés W.G. Cochran (1909-1980):

> En una de sus clases citó sin demostrar un resultado. Tras varios intentos sin que me saliera, le pedí en su despacho si podía hacerme la demostración. Me dijo que en algún sitio la tenía archivada; abrió varios cajones y decidió que era mejor obtenerla de nuevo. Nos sentamos y escribió la misma expresión de la que yo había partido. El camino obvio va en esta dirección, dijo, y escribió una expresión de dos líneas. Ahora supongo que hay que desarrollar esto, y puso una ecuación que ocupaba tres líneas. Miró la expresión y comentó: el único camino parece ser este, y obtuvo una expresión de cuatro líneas y media. Hubo un silencio de unos 45 segundos y dijo, el resultado se debe seguir de esto, escribiendo debajo la expresión que yo le había preguntado. La clase había terminado.

alternativas, para el gen que codifica ese carácter. Fisher demostró matemáticamente que cuanta más variabilidad genética haya en una población, mayor será el ritmo de la evolución. A mayor

variación genética, más cambio evolutivo. Fisher comparaba su teorema con el segundo principio de la termodinámica o ley de la entropía, cuyo incremento es siempre positivo. La selección natural actuaba de manera lenta pero segura, aumentando progresivamente la frecuencia de los genes favorables, que se iban integrando al genoma de la especie, lo que incrementaba la adecuación de los organismos cada vez más. Como consecuencia, la selección tendía a convertir el alelo bien adaptado en el alelo dominante y las mutaciones deletéreas en recesivas.

La genética de poblaciones aportó, empero, solidez matemática a la teoría de la evolución. No obstante, Fisher y Haldane compartieron dos supuestos que fueron muy criticados por Wright. En primer lugar, concibieron la carga genética del individuo como un saco de judías, es decir, como un conjunto de genes que no interactúan entre sí. Fue Wright el que generalizó los modelos simplificados de ambos. En segundo lugar, consideraron las poblaciones al completo, lo que les condujo a visualizar la selección natural como un proceso prácticamente unidireccional, sin ramificaciones. Pero Wright llamó la atención acerca de que las poblaciones grandes generalmente estaban disgregadas en poblaciones locales pequeñas donde triunfaba la endogamia, lo que convertía la selección natural en algo más voluble, dando origen a la noción de *paisaje adaptativo*.

A día de hoy, pese a las encomiables aportaciones de Fisher y el resto de genetistas de la síntesis, siguen existiendo dudas sobre el reparto de papeles que cabe atribuir a la selección natural y las mutaciones en la evolución y, en particular, sobre su acción a nivel molecular. Para algunos, la fuerza evolutiva principal a nivel molecular es simplemente la «deriva genética», es decir, el cambio en las frecuencias alélicas de las especies como consecuencia del efecto estocástico causado por la reproducción (los alelos de los hijos son una muestra aleatoria de los de los padres), primando la presión selectiva a nivel morfológico, a escala de los organismos. Para otros, en cambio, los genes mutantes no son selectivamente neutrales, de forma que el papel de las mutaciones no puede desdeñarse y la selección actuaría tanto a nivel molecular como morfológico. En otras palabras,

no se sabe a ciencia cierta si el sujeto de la evolución es la especie o el genoma. Por otra parte, también hay disenso sobre la continuidad o discontinuidad de los cambios evolutivos (gradualismo). Así, por ejemplo, los partidarios del «equilibrio puntuado» sostienen, frente a los neodarwinianos ortodoxos, que en la evolución se alternan períodos de cambios rápidos con lentos. Nadie discute a Darwin pero los neodarwinistas no presentan un frente único.

ESTADÍSTICA, DARWINISMO Y EUGENESIA

El abanico de motivaciones no estaría completo si no citáramos que Fisher fue un ardiente promotor de la eugenesia, una disciplina que estimuló y guio gran parte de su trabajo en genética humana. Durante sus años en Cambridge, Fisher colaboró con entusiasmo, al igual que otros ilustres científicos (como John Maynard Keynes), con la Eugenics Education Society, fundada en 1907 por Galton y dirigida desde su muerte en 1911 por Leonard Darwin (quien presidió el Primer Congreso Internacional de Eugenesia, celebrado en Londres en 1912 y dedicado a la memoria de Galton). Además, Fisher formó una sociedad eugenésica dentro de los muros de la universidad.

En 1911 ofreció una charla a un grupo de estudiantes simpatizantes en la que expuso la idea de Galton de que la curva normal se aplicaba incluso a las cualidades morales e intelectuales de los hombres, de manera que estos se repartían en varias clases que iban desde los débiles mentales a los genios eminentes. Las virtudes intelectuales y morales constituían, por descontado, rasgos heredables, razón por la cual los matrimonios debían concertarse entre personas de la misma clase. Para Fisher, la obra de Galton *Genio hereditario* era uno de los grandes libros del siglo XIX, comparable a *El origen de las especies* de Darwin, al que en cierto modo completaba.

Uno de los primeros artículos de Fisher vio la luz en 1914 en las páginas de la *Eugenics Review*, la revista estandarte del mo-



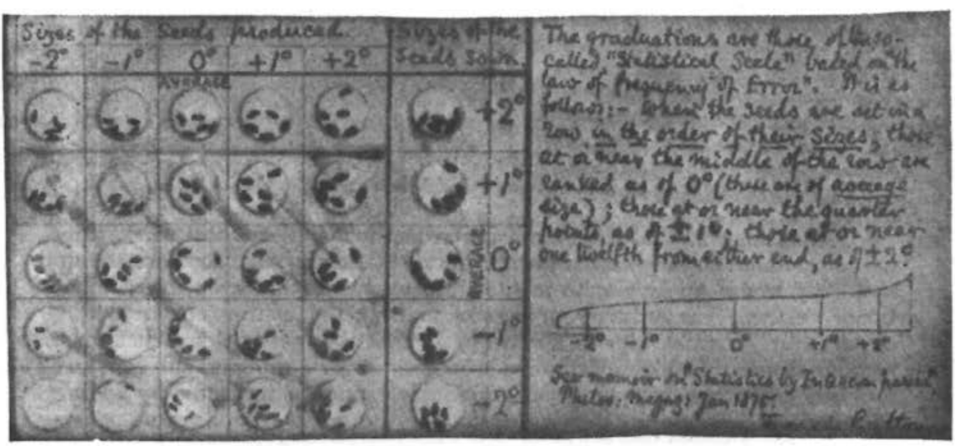


FOTO SUPERIOR: **Una exposición pro eugenésica atrae a una multitud en una feria celebrada en Kansas en 1929.**

FOTO INFERIOR: **Experimentos de Galton sobre la herencia con guisantes. A la derecha se encuentra la representación gráfica de la función de distribución de los resultados, que el científico inglés asemeja —según sus propias palabras manuscritas— a la ley del error.**

UNA CASA EDIFICADA SOBRE ARENA

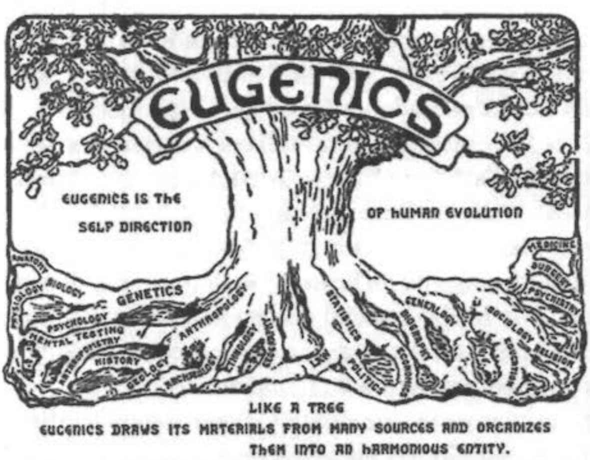
La fuerza motriz del movimiento eugenésico estaba ya en Quetelet, que pensaba que su hombre medio compendiaba las características físicas y morales de una raza. La otra mitad estaba en la idea ligada al evolucionismo biológico de que mediante medidas sociales de selección podían preservarse o alterarse las características raciales (Galton). Sin embargo, los historiadores de la ciencia no se ponen de acuerdo en el peso final que cabe atribuir a la eugenesia en el desarrollo de la estadística. Un bando sostiene que los métodos estadísticos se desarrollaron para resolver los problemas planteados por la investigación en eugenesia. Esta doctrina no solo habría motivado los trabajos de Galton, Karl Pearson o Fisher, sino que habría condicionado su contenido (aunque, por ejemplo, Edgeworth o Yule no compartían el interés por la selección racial). En cambio, el otro bando combate tajantemente esta relación, subrayando que los métodos del laboratorio biométrico del University College eran completamente distintos a los empleados en el laboratorio eugenésico contiguo, o que Karl Pearson nunca se adhirió a la sociedad eugenésica (aunque no lo hizo por su oposición decidida al mendelismo).

Separación definitiva de la estadística y la eugenesia

Probablemente, la biometría y la eugenesia no eran compartimentos estancos. Pero, mientras que ciertos métodos como el test χ^2 encontraron mil y una aplicaciones diferentes (en agronomía, genética, industria, etc.), otros métodos, como los mapas de pedigrí de Galton, no las encontraron. La impronta social de la estadística es innegable: su cristalización se produjo en contacto con la biometría y los intentos por convertir la eugenesia en la reina de las ciencias (como se observa en el cartel del Segundo Congreso Eugenésico Internacional). No obstante, la recurrencia de los métodos estadísticos, es decir, su extensión a una multiplicidad de áreas naturales y sociales, posibilitó su independencia con respecto a la ideología envolvente, a la manera como la mecánica clásica no depende

vimiento eugenésico, donde llegaría a publicar más de 200 artículos entre reseñas de libros y comentarios. Su título era «Algunas esperanzas de un eugenista». El texto, leído previamente para la sociedad universitaria de Cambridge, defendía la eugenesia como vía hacia el progreso de la humanidad. Tres años más tarde, publicó un editorial en que promovía la toma de medidas políticas que incrementaran la tasa de natalidad de las clases profesionales y controlaran la de las clases más bajas. Un tema en el que

hoy de la balística de cañones o la mecánica cuántica de la guerra atómica. La historia de la estadística es una estampa que ilustra a la perfección la imbricación entre historia «interna» y «externa» de la ciencia: la eugenesia fue la pasarela que permitió conectar la biología evolutiva con la estadística y, de resultas de ello, impulsar la creación de las primeras instituciones estadísticas modernas. Fruto de esta sinergia, la estadística cobró fuerza suficiente para arraigar en otros ámbitos científicos, lo que dictó su independencia respecto de la ideología eugenésica y, de paso, la devolución del favor prestado por la biología evolutiva con creces, al determinar el cierre de la genética de poblaciones y el establecimiento de la teoría sintética de la evolución.



«La eugenesia es la autodirección de la evolución humana.» Ese era el lema del Segundo Congreso Internacional de Eugenesia (1921), que representaba a la eugenesia como el árbol que unifica la diversidad de disciplinas humanas y sociales, con la genética y la estadística en una posición preeminente.

se explayó en los últimos capítulos de *La teoría genética de la selección natural.*

Fisher atribuía el declive de las civilizaciones al hecho de que se alcanzaba un momento histórico en el que la fertilidad de las clases altas comenzaba a decaer en detrimento de la de las clases bajas (las peor adaptadas, a su entender, atendiendo a las cualidades mentales). Utilizando datos extraídos de los censos de Gran Bretaña, Fisher mostraba la relación in-

versa entre fertilidad y estatus social: las clases altas tenían una baja fertilidad, y las bajas, una tasa alta de fertilidad. Las familias con un alto estatus social no podían permitirse dejar mucha descendencia, ya que tener un número reducido de hijos era una ventaja económica. Para superar esta lacra, el eugenista británico proponía que por medio de subsidios estatales se paliara el gasto excesivo que suponía tener una prole abundante. Quizá Fisher, que tuvo dos hijos y seis hijas, estaba expresando aquí una vivencia personal.

Coincidiendo con la publicación del libro en 1930, Fisher dedicó bastante tiempo a colaborar con la sociedad eugenésica abanderada por Leonard Darwin. Así, al Tercer Congreso Internacional de Eugenesia, celebrado en Nueva York en 1932, acudió para hablar en lugar de su mentor, dada su avanzada edad. Todavía más: Fisher participó muy activamente en la campaña emprendida por la sociedad a favor de la aprobación de una ley que permitiese la esterilización en base a criterios eugenésicos. A diferencia de Estados Unidos, Alemania, Dinamarca y otros países protestantes, en Gran Bretaña no se logró la adopción de leyes de esterilización voluntaria ni forzosa. No obstante, debe matizarse que los eugenistas británicos siempre incidieron más en la repercusión de la clase social que en la de la raza natural sobre la herencia de las cualidades mentales, en contraposición de sus homólogos norteamericanos o alemanes.

Tras su mudanza al University College desde Rothamsted en 1933, Fisher prosiguió los estudios eugenésicos en el Laboratorio Galton. Junto con otros colegas, profundizó en la recolección de datos sobre pedigrís humanos, así como en el estudio de los grupos sanguíneos y el factor Rhesus. Y en 1950 se opuso frontalmente a la Declaración sobre la Raza de la Unesco, que sostenía que este concepto era una mera herramienta clasificadora, disociada de las culturas, las etnias o las puntuaciones en los test de inteligencia. Fisher mantenía que la experiencia de cada día mostraba que las diferencias innatas intelectuales y emocionales entre razas no podían minimizarse.

En el presente, la palabra *eugenesia* posee un sabor rancio, pasado de moda. Lo que fue una idea fuerza, parece inerte. Sin

embargo, con el propósito de contextualizar la creencia de Fisher en las virtudes de la eugenesia, hay que apuntar que a día de hoy muchos científicos y personas en general se muestran partidarios de la ingeniería genética, aplicada no solo a patologías, sino a rasgos físicos seleccionables, como el color del pelo o de los ojos del neonato.

CAPÍTULO 5

A vueltas con la inducción y el método científico

Paralelamente a sus descubrimientos matemáticos y biológicos, Fisher dedicó parte de su tiempo a meditar sobre el significado de la probabilidad y el alcance de los métodos estadísticos, en especial de la inferencia bayesiana en comparación con la inferencia frecuentista, que defendía como más adecuada. No hubo costura del tejido estadístico que Fisher no repasara, lo que le condujo a polemizar con Jerzy Neyman y Egon S. Pearson a propósito de los contrastes de hipótesis y, ya en sus últimos años de vida, con los médicos a colación del tabaco y el cáncer.

Después de atravesar una larga crisis económica y anímica, Fisher regresó en 1943 a Cambridge, su *alma mater*, para ocupar la cátedra de Genética, sucediendo a R.C. Punnett. La convivencia con Fisher no era fácil, dada su personalidad contradictoria: lúcido y ofuscado, feroz y amistoso, avaro y espléndido. Todo a la vez. A los apuros monetarios se sumaba el duro trabajo, así como el cuidado de la prole. La desatención al estado de salud de su esposa condujo a una crisis doméstica irreversible en 1942. Además, ese mismo año, el mayor de sus hijos varones, que se había alistado como piloto de combate en la Segunda Guerra Mundial, falleció en un accidente aéreo sobre Sicilia, lo que dejó a ambos cónyuges destrozados. El matrimonio se rompió cuando Fisher se trasladó a Cambridge... solo.

La estadística matemática desarrollada por Fisher durante la década de los felices años veinte en seguida sembró controversia (personal y conceptual). Esta circunstancia motivó que Fisher reflexionara profundamente sobre la lógica intrínseca de los nuevos métodos de inferencia científica, la inferencia estadística denominada hoy día *clásica*. Ya en 1935 publicó un artículo tentativo sobre el tema bajo el título «La lógica de la inferencia inductiva», cuya lectura en la Real Sociedad de Estadística a finales del año anterior había suscitado mil y una réplicas. Pero sería en la década de 1950 cuando más páginas dedicara a la cuestión. Al

polémico artículo «Métodos estadísticos e inducción científica», presentado a la Real Sociedad de Estadística en 1955, le siguió el libro *Métodos estadísticos e inferencia científica*, un mamotreto publicado en 1956 donde Fisher ahondaba en los aspectos más filosóficos de la inferencia estadística.

En esta última obra, Fisher intentó ofrecer una perspectiva unificada de la inferencia, englobando sus tres aproximaciones en vida al problema: el método de máxima verosimilitud, los test de significación y la probabilidad fiduciaria (cuya definición se explicará más abajo). El libro tomó la forma de un repaso de la inferencia estadística desde Bayes al presente. Por el camino, Fisher condenaba a la hoguera a Bayes y a Karl Pearson, entre otros «falsos profetas». El estadístico británico aprovechó además la ocasión para mostrar su animadversión y desdén para con los estadísticos estadounidenses, cuya concepción de la estadística presumía que era puramente matemática, sin contacto alguno con las ciencias experimentales. Para algunos colegas, como Maurice Kendall, este libro —como el panfleto de 47 páginas sobre el cáncer y el hábito de fumar que vio la luz en 1959— nunca debería haber sido escrito.

Sea como fuere, son tres los puntos de fricción a los que Fisher prestó atención: el significado de la probabilidad, las carencias de la inferencia bayesiana y la lógica de los contrastes de hipótesis.

DEFINIR LA PROBABILIDAD

A pesar de que la palabra *probabilidad* era de uso corriente en las lenguas emparentadas con el latín (donde *probable* significaba algo así como «merecedor de aprobación»), el concepto matemático de probabilidad no hizo su entrada, como dijimos en el primer capítulo, hasta alrededor de 1660. Y lo hizo arrastrando, desde su nacimiento, una singular dualidad. La idea emergió como un Jano bifronte que representaba una mutación de la idea renacentista de los signos. Una afirmación era probable cuando estaba bien atestiguada. Con el Renacimiento, el mundo comenzó a testificar por sus signos. No solo los libros de los doctores constituían un

testimonio válido. Ahora también lo era, por decirlo con Galileo, el libro de la naturaleza. De modo que el signo probable era una señal frecuente, repetida, mediante la cual el mundo daba testimonio, credibilidad (del mismo modo que el humo es un signo del fuego).

Por tanto, la probabilidad surgió ligada, por un lado, a la creencia y, por otro, a la frecuencia. Al igual que el modo escolástico de la posibilidad, la probabilidad podía predicarse *de dicto* (acerca de las proposiciones y su evidencia) o *de re* (acerca de las cosas y de la tendencia, exhibida por algunos dispositivos de azar, a producir frecuencias relativas estables). La palabra *probabilidad* fue usada por primera vez para denotar algo medible en la *Lógica* de Port-Royal, un manual sobre el arte de pensar impreso en torno a 1662 por varios colaboradores de Pascal afincados en ese enclave jansenista.

Tanto Poisson, en su obra de 1837 sobre la ley de los grandes números, como Cournot, en su libro de ciencia moral publicado en 1843, aclaraban que la probabilidad mezclaba dos nociones que había que distinguir con precisión de cirujano: por una parte, la *chance* o probabilidad física, que cuantificaba la facilidad o propensión —como se dice actualmente— a aparecer que tiene un suceso; por otra, la *probabilité* o probabilidad epistémica, que medía la credibilidad que merecía la ocurrencia del suceso. Mientras que la primera aludía a una propiedad objetiva del suceso (la posibilidad de que ocurra, muy útil para modelar), la segunda era subjetiva (relativa a nuestro conocimiento, de utilidad al inferir).

Curiosamente, un siglo antes, el reverendo Thomas Bayes había dejado escrito: «por *chance* entiendo lo mismo que *probabilidad*». Pero a la altura de 1850, el mundo ya no era como en la época de Bayes y Laplace. El aspecto objetivo de la probabilidad pasó a ser mucho más determinante que el subjetivo, sencillamente porque el mundo rebosaba de frecuencias. El alud de números impresos inclinó la balanza. De hecho, John Venn, en la *Lógica del azar* (1866), apostó por un enfoque frecuencial más que personal de la probabilidad.

Sin embargo, la inferencia estadística decimonónica siguió siendo claramente bayesiana (para estimar incertidumbres se usa-

SOLUCIONES AXIOMÁTICAS

Las dos interpretaciones de la probabilidad comparten un mismo formalismo matemático: los axiomas de Kolmogórov (1903-1987), formulados por el matemático soviético en 1933. Cualquier interpretación de la probabilidad que satisfaga estos axiomas —y hay más— es una buena realización del concepto. Los axiomas propuestos respetaban las intuiciones plasmadas en la definición clásica (la «regla de Laplace», solo aplicable a casos equiprobables) y en la definición frecuentista (el teorema de Bernoulli, solo aplicable a fenómenos susceptibles de repetirse) de la probabilidad, además de conectar la teoría de la probabilidad con la teoría de conjuntos y la teoría de la medida, transformándola en una teoría matemática firme que en seguida se difundió por Centroeuropa permitiendo la prueba de múltiples teoremas. Por su parte, la interpretación subjetiva de la probabilidad (como grado de creencia en una proposición o de adhesión a la verificabilidad de un suceso, variable en cada persona, aunque sujeta a reglas bastante estrictas de coherencia interna) fue formalizada independientemente por el estadístico italiano Bruno de Finetti (1906-1985) en 1937 y difundida por Leonard J. Savage (1917-1971) en 1954, quien resucitó la inferencia bayesiana y recuperó este enfoque de la probabilidad relacionado con la utilidad (noción introducida por Daniel Bernoulli, sobrino de Jakob, en 1737 y más tarde por Frank P. Ramsey en 1931).


Andréi Nikoláyevich Kolmogórov.

ban los métodos inversos de probabilidad de Bayes y Laplace). Solo cuando la sobrepoblación de números, de frecuencias registradas accesibles, fue un hecho más allá del campo astronómico (acúmulo de datos entresacados de la sociología, la biología o la agronomía), pudo desarrollarse —gracias a Fisher, como vimos en el capítulo 3— la inferencia estadística objetiva en detrimento de la bayesiana o subjetiva. Con la observación continuada de regularidades en otras áreas naturales distintas de la bóveda celeste, la in-

terpretación subjetiva de la probabilidad como grado de creencia, de estirpe laplaciana, quedó definitivamente marginada por la interpretación objetiva o frecuentista: las probabilidades ya no se basarían en creencias sino en frecuencias empíricas. Desde el principio Fisher fue consciente de que cada interpretación de la probabilidad apuntaba a una teoría distinta de la inferencia, ya que los conceptos probabilísticos son los ladrillos de la inferencia estadística.

«ALL YOU NEED IS BAYES...»

Para muchos científicos, la estadística tiene la responsabilidad de responder una pregunta fundamental: ¿cuándo es correcto afirmar que un conjunto de observaciones aporta evidencia a favor o en contra de una hipótesis? El recurso más antiguo para dirimir esta cuestión se remonta a 1763: el teorema de Bayes, aparecido en el *Ensayo hacia la solución de un problema en la doctrina del azar*, firmado por el reverendo Thomas Bayes. Este teorema, precursor de los métodos inversos de probabilidad y de la inferencia bayesiana, era el resultado central de un ensayo destinado en espíritu a combatir la crítica escéptica a la inducción planteada por el filósofo escocés David Hume en *Sobre los milagros*, ya que ofrecía una discusión matemática del incremento de probabilidad entendida como credibilidad.

Solo dentro de este contexto teológico influido por Newton puede entenderse que, por ejemplo, el doctor John Arbuthnot, concupiscente médico de la corte aficionado a calcular probabilidades como la de que una mujer de veinte años conservara su virginidad o un joven hubiera sido infectado de gonorrea, realizara en 1710 la que pasa por ser la primera prueba de significación de una hipótesis estadística: si la posibilidad de nacimiento de un varón fuese igual a la de una hembra (esto es, 1/2), la probabilidad de que se registrasen —como se había constatado— ochenta y dos años consecutivos en que nacían más hombres que mujeres sería de $(1/2)^{82}$, o sea, prácticamente cero. Por ende, la hipótesis de igualdad de sexos al nacer debía ser rechazada, y Arbuthnot

interpretaba esta regularidad como un argumento (inductivo) a favor de la divina providencia. En esta línea, la fórmula de Bayes permitía emitir juicios probabilísticos sobre la validez de una hipótesis (probabilidad *a posteriori*) basándose en los datos (verosimilitudes), pero también en la apreciación subjetiva que la hipótesis mereciese (probabilidad *a priori*).

«Las causas que llevaron a Bayes a su teorema eran más teológicas y sociológicas que puramente matemáticas.»

— **Karl Pearson (1926).**

No obstante, el problema de la probabilidad inversa había cobrado forma con la contribución de Jakob Bernoulli en 1713. El matemático suizo le había comunicado por carta a Leibniz en 1704 que había encontrado un teorema que le permitía calcular *a posteriori*, con una aproximación determinada, las probabilidades desconocidas de los sucesos conocidos empíricamente tan bien como si aquellas le fuesen conocidas *a priori*, de entrada. Sin embargo, como explicamos en el primer capítulo, el teorema áureo de Bernoulli no era exactamente un ejemplo de probabilidad inversa, porque lo que el teorema venía a afirmar es que, «conocida» la probabilidad de ocurrencia de un suceso, la frecuencia relativa con que este suceso ocurre tiende a ese número (ley débil de los grandes números). En cuanto tal se trata de un teorema puro e incuestionable de la teoría de probabilidades. Así, Bernoulli fue capaz de deducir el número de veces que hay que lanzar un dado simétrico (legal) para que, con «certeza moral» (esto es, con probabilidad mayor o igual que 0,999, un estándar análogo al que los estadísticos modernos usan hoy del 95% o 99% de confianza), la frecuencia relativa con que salga el 6 difiera de $p = 1/6$ (su probabilidad, que, nótese, se supone conocida) en no más de 0,01: 1 388 889 veces. En el teorema la probabilidad p estaba fija y se calculaba la probabilidad de observar ciertos datos, sabiendo que la frecuencia relativa de éxitos f_n tendía a p cuando el número de experimentos n aumentaba. Bernoulli hacía aseveraciones acerca de lo que en la época se llamaban *problemas directos de probabi-*

lidad, problemas en los que se suponía conocida la probabilidad de éxito y se calculaba la probabilidad de cualquier sucesión de éxitos y fracasos.

Pero si no se conocía p, ¿podía usarse todavía el teorema? Paradójicamente, Bernoulli introdujo su teorema precisamente para aquellos casos en los que no se tenía conocimiento previo de p. Sin embargo, resistió la tentación de invertir el teorema, conformándose con acotar los posibles valores de p entre dos límites (anacrónicamente, diríamos que realizó una estimación por intervalo de p para un cierto nivel de confianza, con certeza moral; un procedimiento que tendría continuación con la teoría astronómica de los errores probables, que construiría estimaciones por intervalo con un nivel de confianza del 50%). En otras palabras, Bernoulli descubrió cómo computar la siguiente probabilidad (donde se conoce p): $P(p$ está en $f_n \pm \varepsilon \mid p)$. Y le habría resultado tentador tomar los valores calculados aquí como los valores de la probabilidad $P(p$ está en $f_n \pm \varepsilon \mid f_n)$, donde se ha sustituido el conocimiento de p por el de f_n. Naturalmente, este paso es falaz, pues la segunda expresión no se deduce de la primera. Parece que fue Laplace quien sucumbió a la tentación de «invertir» el teorema, e inferir la probabilidad p a partir de la frecuencia observada f_n, a pesar de que esta tendencia ya estaba en el propio Bernoulli, quien de haber tenido éxito en su empeño habría resuelto el problema de la inducción, de ascender de lo particular a lo general, de la muestra a la población (la inferencia inductiva).

La solución completa de Laplace a este problema pasó, canonizando la interpretación epistémica de la probabilidad, por el teorema de Bayes, que considera la probabilidad desconocida p como una variable aleatoria. El opúsculo de Bayes fue el primer intento sistemático de calcular la segunda probabilidad antes expresada: mediante una asignación *a priori* de probabilidades y por medio de la fórmula de Bayes, se calculaba la probabilidad pedida. Presuponiendo una distribución *a priori* de p sobre el intervalo [0,1], Laplace calculó a partir de los datos disponibles la probabilidad (*a posteriori*) de que p estuviese a menos de una cierta distancia ε de la frecuencia relativa f_n observada. Dado el número de veces que había salido 6, calculaba la probabilidad de

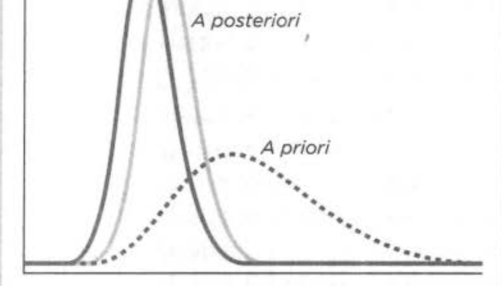
La distribución *a priori* y la verosimilitud aparecen, respectivamente, con línea entrecortada y con línea continua negra. La distribución *a posteriori*, calculada mediante el teorema de Bayes, se representa con línea continua de color gris (en el eje horizontal se colocarían los posibles valores del parámetro θ que se desea estimar). Como puede observarse, la distribución *a posteriori* se encuentra entre medias, a medio camino de la distribución *a priori* y la verosimilitud. De hecho, en este ejemplo, se parece mucho más a la verosimilitud que a la *a priori*, lo que muestra cuánto hemos aprendido de los datos.

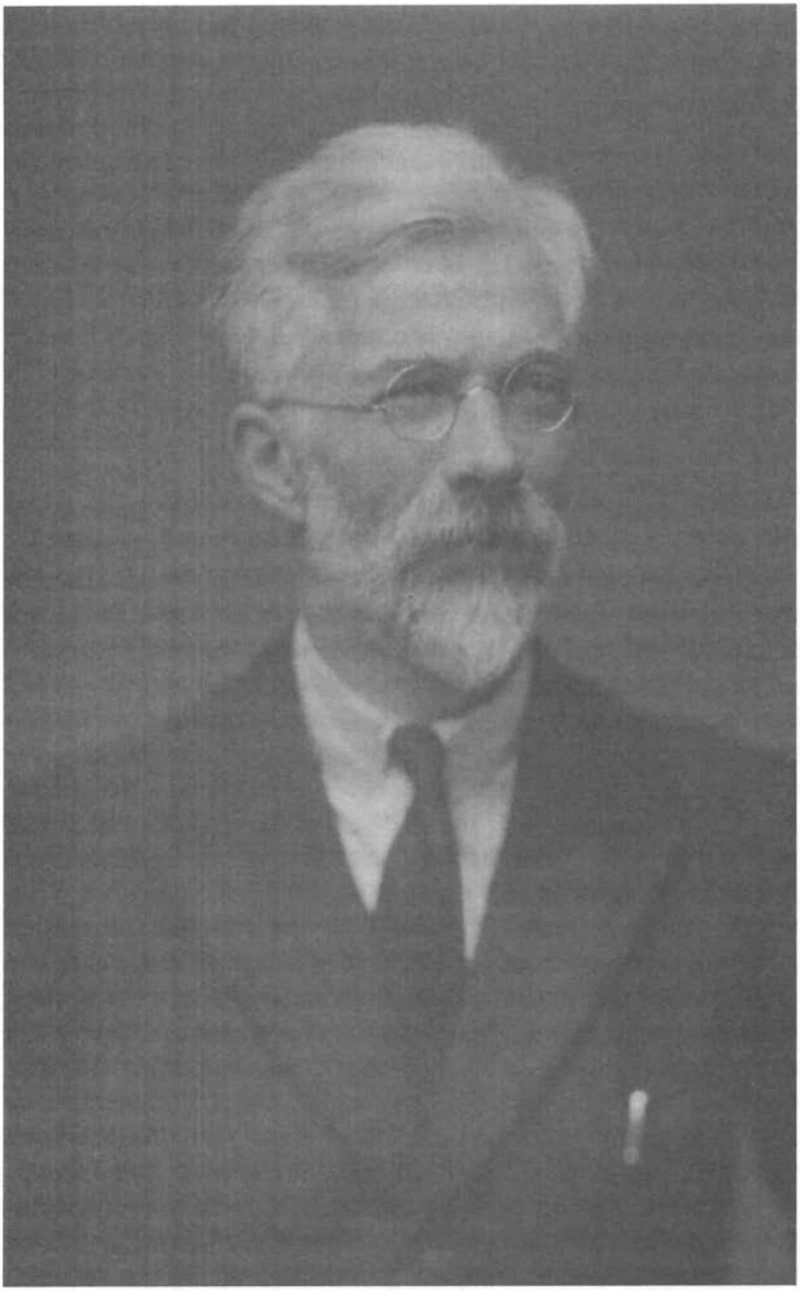
que la probabilidad de salir 6 estuviese en un entorno de la frecuencia relativa observada.

Los estadísticos bayesianos buscan conocer la probabilidad de que cierto parámetro desconocido θ se encuentre entre dos valores prefijados. Para ello necesitan dos cosas: en primer lugar, las verosimilitudes $P(X|\theta)$, es decir, las probabilidades de observar la muestra extraída de la población dependiendo del valor que tome el parámetro; y, en segundo lugar, la probabilidad *a priori* de θ o distribución *prior* de θ, que mide la probabilidad de que el parámetro desconocido se encuentre entre dos límites cualesquiera. La distribución *a posteriori* $P(\theta|X)$, calculada mediante la regla de Bayes, no es sino un compromiso entre la distribución *a priori* y la verosimilitud, entre lo que sabíamos y lo que hemos aprendido de los datos observados (figura 1).

La preferencia del siglo XIX por los números y la objetividad incentivó a los matemáticos a buscar alternativas a un procedimiento que era mirado con suspicacia. Fisher hizo de la lucha contra la inferencia bayesiana una de las razones de su vida científica. A su entender, los métodos estadísticos habían conducido a una comprensión más completa de la lógica inductiva, constituyendo la base de la inferencia científica, pues la inferencia inductiva era, a diferencia de la deductiva, ampliadora del conocimiento (porque permite aprender de la experiencia, aunque siempre con un cierto grado de incertidumbre, pero que al poder cuantificarse hace la inferencia perfectamente rigurosa). Ahora bien, mientras que el papel principal en la inferencia deductiva o directa (de lo general a lo particular, de la población a la muestra) lo tomaba la probabilidad, la inferencia inductiva o inversa (de lo particular a lo general, de la muestra a la población) estaba reservada a la verosimilitud y, en algunos casos, a la probabilidad fiducial. Bajo ningún concepto a la probabilidad bayesiana.

FIG. 1





Ronald A. Fisher en 1943, año en que volvió a la Universidad de Cambridge para ocupar la cátedra de Genética, tras atravesar graves problemas familiares que acabaron con la disgregación de su matrimonio.

Entre otras endebleces, Fisher criticaba que los bayesianos transformaban clandestinamente la inferencia inversa o inductiva en una inferencia directa, en una deducción probabilística, al postular un conocimiento de partida: la distribución *a priori* del parámetro θ. En cuanto ecuación matemática, la fórmula de Bayes podía ser indiscutible (aunque, para Fisher, era poco o nada evidente), pero su empleo requería asignar una probabilidad *a priori* a la verdad de la hipótesis que se valora, un número borroso sujeto a discusión. No era plausible que en situaciones de completa ignorancia, uno admitiera que debe asignar a todos los posibles valores de θ la misma probabilidad (distribución uniforme) o una probabilidad que depende del estado de información en que se encuentre cada uno (probabilidad subjetiva), de manera que dos investigadores pueden usar dos priores inconsistentes entre sí cayendo en el subjetivismo más inaceptable. (De hecho, actualmente se conocen algunas paradojas, como la «paradoja de Lindley», que muestran cómo la inferencia bayesiana puede fallar estrepitosamente si se eligen priores inadecuadas: toda la probabilidad se deposita *a posteriori* en ciertos valores del parámetro se observe lo que se observe.) Además, el hecho de que con el aumento del tamaño muestral la forma precisa de la distribución prior perdiera relevancia en relación con la verosimilitud (como en el gráfico que antes hemos mostrado en la figura 1, pág. 134), llevaba a Fisher a afirmar que lo más natural era extraer conclusiones sin suposiciones *a priori* de ninguna clase.

No obstante, para Fisher la inferencia inductiva era posible aunque no transcurriera por los canales bayesianos. A diferencia del filósofo Karl Popper, Fisher no creía que la ciencia debiera retornar a un simple modelo demostrativo, alejado de la práctica experimental. La mayoría de matemáticos, demasiado entrenados en el arte de la deducción, confundían una inferencia incierta (donde la incertidumbre es cuantificable) con una inferencia no rigurosa. El aprendizaje de la experiencia se producía por medio de los test de significación, que, como reflejamos en el tercer capítulo, servían para extraer conclusiones de los datos observados sin referencia alguna a creencias previas (*a priori*). Y la verosimi-

litud era la medida de creencia racional; porque, a diferencia de la probabilidad (que solo permite razonamientos deductivos, pues la fórmula de Bayes ya parte de la prior), posibilita razonamientos inductivos, al ser lo que se evalúa en los test.

> «Tiene un error lógico en la primera página que invalida las restantes 395, y es que adopta el postulado de Bayes.»
>
> — Fisher sobre el libro *Teoría de la probabilidad* (1939) del astrónomo Harold Jeffreys.

En torno a 1930, Fisher encontró que, en ciertas situaciones especiales, era factible transformar los conocimientos logrados sobre el parámetro en sentencias probabilísticas sin usar el teorema de Bayes. A través de un oscuro argumento, Fisher definía una distribución de probabilidad sobre el parámetro θ en base a los datos y sin tomar en cuenta ninguna distribución *a priori*. Era la denominada *probabilidad fiducial*. Fisher pensaba en $P(X|\theta)$ como una función en dos variables y, cuando sustituía el valor muestral observado X y podía despejar adecuadamente θ en función de X, explotaba la consideración de $P(\theta|X)$ como una distribución de probabilidad en θ a efectos prácticos. Había encontrado un método para invertir afirmaciones probabilísticas sobre las observaciones una vez dado el valor del parámetro en afirmaciones probabilísticas sobre el parámetro a partir de las observaciones.

En el argumento fiducial hay una transmisión de probabilidad de X a θ, del estadístico muestral al parámetro, que es intuitiva pero confusa; porque cambia el estatus del parámetro, que pasa de ser un valor desconocido pero constante a ser una variable aleatoria. Para Leonard J. Savage, «la aproximación fiducial de Fisher era un intento de hacer una tortilla bayesiana sin romper ningún huevo bayesiano», ya que lo único que diferenciaba al método fiducial del método de Bayes era la ausencia de conocimiento *a priori*. De hecho, la distribución fiducial podía calcularse como una distribución *a posteriori* respecto de una prior no informativa (neutra, uniforme). Esto provocó que Fisher suavizara su posición, de manera que en su libro de 1956 se muestra partidario de la aproximación

¿SALDRÁ EL SOL MAÑANA?



Retrato idealizado de Thomas Bayes.

Persiguiendo refutar a Hume, quien había escrito que únicamente era probable que el Sol saliera de nuevo al día siguiente, Richard Price (1723-1791), el filósofo que se encargó de publicar póstumamente el legajo de Bayes, empleó el teorema de su colega para calcular la probabilidad de que el Sol así lo hiciera. Teniendo en cuenta el número de días que había venido amaneciendo ininterrumpidamente, Laplace mejoró los cálculos alcanzando la «regla de sucesión»: si un hecho se repite seguidamente cualquier cantidad de veces, la probabilidad de que ocurra una vez más es igual a este número más 1 y dividido por este mismo número más 2. Así, si suponemos que el Sol ha salido invariablemente durante 5000 años, o sea, 1826213 días (Laplace pensaba que la Tierra era muy joven y le adjudicaba solo 5000 años de existencia), la probabilidad de que salga mañana es de 1826214/1826215 (≈ 99,9999%). No obstante, como buen astrónomo, Laplace subrayaba que en el caso de este tema se trataba más bien de un problema de mecánica celeste que de probabilidad; porque, por esta regla, cuanto mayores nos vayamos haciendo, mayor resultará la probabilidad de vivir más. De modo que una persona de ochenta años tendrá mayor probabilidad de vivir un día más que una de solo veinte años. Lo que carece de sentido.

bayesiana cuando la información muestral sobre el parámetro sea lo suficientemente extensa, ya que en el cálculo de la distribución *a posteriori* mediante el teorema de Bayes la verosimilitud será determinante (como en el gráfico visto en la figura 1, pág. 134). En otro caso, era partidario del argumento fiducial.

Los esfuerzos por suplantar el teorema de Bayes, encarnados en personalidades tan importantes como Fisher, no lo consiguieron, y a lo largo de la segunda mitad del siglo XX se ha asistido a un resurgir de la inferencia bayesiana, el enfoque ciertamente más antiguo dentro de la inferencia estadística, en conexión con la teoría de la decisión. El bayesianismo intenta ser una aproximación

formal, algorítmica, a esa vaga idea que sería «aprender de la experiencia para decidir mejor». Da un procedimiento para combinar nuestra información *a priori* con la muestra a fin de obtener una inferencia que tenga en cuenta toda la información disponible.

FIG. 2

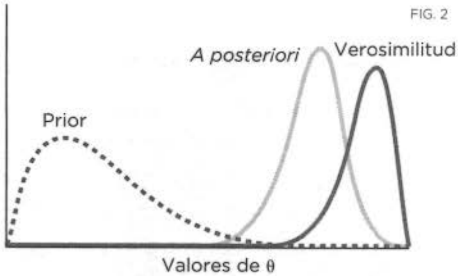


FIG. 3

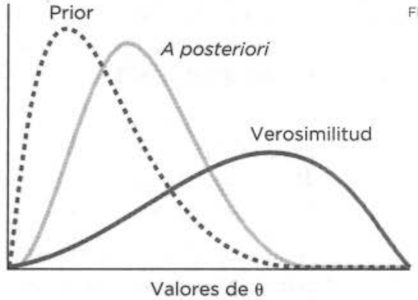
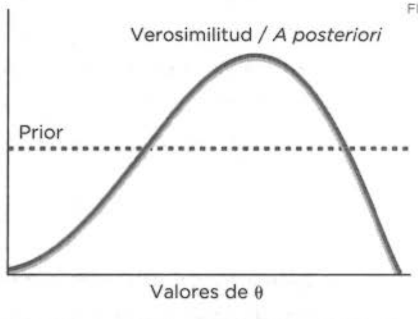


FIG. 4



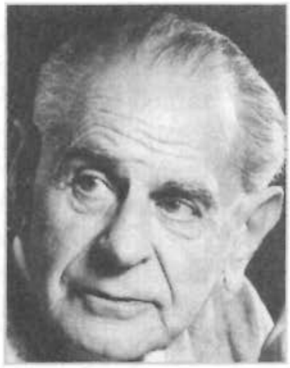
A día de hoy algunos estadísticos sostienen que la inferencia del futuro será bayesiana o no será, ya que los métodos clásicos fallan en ocasiones en su precisión, no toman en cuenta la información proveniente de estudios previos y tampoco ayudan a valorar la credibilidad de una hipótesis. Mientras que la inferencia clásica supone que el parámetro θ está fijo y pretende estimarlo, la inferencia bayesiana lo interpreta como una variable aleatoria de modo que la probabilidad $P(\theta|X)$ es objeto de estudio. Si el tamaño de la muestra X es grande, ambos métodos ofrecen en general los mismos resultados, ya que la información muestral pesa mucho más que la información *a priori* (como puede observarse en la figura 2, la distribución *a posteriori* se asemeja más a la verosimilitud que a la prior). Pero si la muestra es pequeña, ambos métodos pueden conducir a resultados distintos, ya que la información *a priori* pesa entonces más que la muestral (en la figura 3 la distribución *a posteriori* se diferencia bastante de la verosimilitud). Sin embargo, en situaciones de máxima incertidumbre, tomar como distribución inicial una distribución neutra (no informativa, uniforme) recupera los resultados clásicos (en la figura 4

la distribución *a posteriori* y la verosimilitud coinciden porque la prior es uniforme). No obstante, los métodos bayesianos a veces son difíciles de aplicar, necesitando del cálculo numérico y del método de Monte-Carlo. Quizá su repunte en la actualidad sea indisociable de la extensión del ordenador.

Frente al bayesianismo subjetivo, se reivindica hoy un bayesianismo objetivo, en el que las probabilidades *a priori* no están basadas en las creencias personales previas del estadístico, sino en ciertas distribuciones iniciales de referencia, regladas. Algunos estadísticos sostienen que esta vía es la mejor ruta para unificar las inferencias bayesiana y clásica. De hecho, tanto Bayes como Laplace empleaban priores objetivas: distribuciones uniformes. Sin embargo, los bayesianos ortodoxos consideran este bayesianismo como deshonesto, y reclaman, con De Finetti o Savage, el empleo de probabilidades personales, confiando en el poder de la evidencia empírica para neutralizar las diferencias en las asignaciones de probabilidad inicial de distintos sujetos, sin que haga falta introducir otras constricciones que la consistencia o coherencia con los axiomas de la teoría matemática de la probabilidad. La traba es que si una persona piensa que cierta hipótesis es imposible, mientras que otra le asigna cierta probabilidad *a priori* positiva, el teorema de Bayes nunca será capaz de ponerlas de acuerdo pese a toda la evidencia que se reúna.

Obviamente, los bayesianos objetivos tratan de neutralizar este relativismo inicial (que los subjetivos salvan fiando a un hipotético límite futuro común) constriñendo la asignación de probabilidades iniciales mediante diversas reglas, como el «principio de razón insuficiente» de Laplace (o de *indiferencia*, según lo rebautizó el economista John Maynard Keynes), que asigna la misma probabilidad a todos los sucesos desconocidos. Ahora bien, si para ser objetivos se usan siempre distribuciones uniformes o cuasi-uniformes, el estadístico bayesiano solo recupera los resultados del estadístico clásico, porque para poder superarle —exhibiendo, por ejemplo, estimaciones de un parámetro con menor error—, ha de introducir en general una distribución *a priori* distinta, en cuyo caso el debate entre estadísticos clásicos y bayesianos vuelve al punto de partida.

CUANDO KUHN CONOCIÓ A BAYES



Sir Karl Popper.

La revitalización de los métodos bayesianos ha tenido mucho que ver con las corrientes en boga en el ámbito de la filosofía de la ciencia. Los filósofos de la ciencia distinguen dos clases de razonamiento no deductivo. Por un lado, está la «inducción» o inferencia bajo incertidumbre y, por el otro, la «abducción» o creación especulativa de hipótesis teóricas para explicar los fenómenos. Tanto la inducción como la abducción han intentado recibir un tratamiento probabilístico por parte de los epistemólogos atravesado el ecuador del siglo XX. La primera muesca se debió a Rudolf Carnap, un filósofo perteneciente al Círculo de Viena que terminó afincado en Estados Unidos, y que pretendió suturar la herida de muerte de la lógica inductiva: el hecho de que la seguridad del razonamiento inductivo palidece al compararla con la del deductivo. Para ello, planteó una teoría axiomática de la confirmación basada en una serie de reglas que buscaban cuantificar la probabilidad inductiva o lógica de una hipótesis, es decir, la probabilidad de una hipótesis H a partir de la evidencia e disponible. Si $P(H|e)=1$, quería decirse que e implicaba H. En cambio, si $P(H|e)=0$, e implicaba la negación de H. Finalmente, si $0<P(H|e)<1$, este número medía el grado en que la estructura lógica de e implicaba parcialmente H. Esta formulación retomaba una idea que ya estaba en los tratados que escribieran Keynes y el astrónomo bayesiano Harold Jeffreys, para los que toda probabilidad inductiva era en el fondo condicional, relativa a la evidencia accesible. En suma, para Carnap, confirmar inductivamente era igual que implicar deductivamente, pero su lógica en seguida se reveló como lastrada por graves problemas técnicos y conceptuales.

La verosimilitud de Popper

Karl Popper, en concreto, azotó furibundamente este inductivismo, estableciendo una larga polémica. Al igual que Fisher, rechazaba tajantemente el uso inductivo de la probabilidad, proponiendo el concepto de *verosimilitud* como sustituto (aunque la verosimilitud popperiana no se define igual que la verosimilitud fisheriana). A todos los efectos, Popper fue a los filósofos inductivistas lo que Fisher fue a los estadísticos bayesianos. El empeño de algunos filósofos por definir una lógica probabilística apropiada para las teorías e hipótesis ha fracasado; pero el reconocimiento de que la ciencia envuelve juicios y valoraciones subjetivas, como puso de manifiesto Thomas Samuel Kuhn en su obra *La estructura de las revoluciones científicas* (1962), ha puesto las esperanzas de muchos epistemólogos en la inferencia bayesiana.

INDUCCIÓN, DEDUCCIÓN Y DECISIÓN

La escuela bayesiana no fue la única a la que se enfrentó Fisher. Dentro de la inferencia objetiva auspiciada por el estadístico británico creció otra escuela en torno a las aportaciones de Egon Pearson y, en especial, Jerzy Neyman (1894-1981). Este matemático de origen polaco se interesó de joven por la aplicación de la estadística en agricultura. Gracias a una beca, pasó el año académico de 1925-1926 en el laboratorio de Karl Pearson, aunque se desilusionó al descubrir que el gigante inglés ignoraba la matemática abstracta continental. El siguiente curso académico optó por pasarlo en París, asistiendo a las clases de Henri Léon Lebesgue. Si no hubiese sido por el fructífero contacto epistolar con Egon Pearson, Neyman hubiera cambiado la estadística por las integrales a su vuelta a Varsovia.

Cuando Karl Pearson cedió el testigo a su hijo Egon, este no tardó en invitar a Neyman al University College. Juntos formaron un tándem que concibió un nuevo paradigma estadístico a partir de los test de significación elaborados por Fisher: los contrastes de hipótesis, cuyo planteamiento perfeccionaron en varios artículos espaciados entre 1928 y 1933, cuando dieron a conocer el lema fundamental que juega un papel crucial en la teoría. Al año siguiente, Neyman reformuló la estadística inductiva al asentar la estimación mediante intervalos de confianza —que en cierto sentido mejoraban los intervalos fiduciales de Fisher— y al dar inicio a la teoría moderna del muestreo: el muestreo aleatorio, en sus diferentes modalidades, como principio básico de aplicación de la estadística.

Al comienzo, Fisher calificó el trabajo de Neyman de luminoso y celebró que plantease la inferencia en términos no bayesianos (la lectura del tratado de probabilidad escrito por Richard von Mises le había convertido en un frecuentista radical). Pero, coincidiendo con el ingreso de Neyman en la Real Sociedad Estadística en 1935, Fisher rompió dramáticamente toda relación con él, al atacar su investigación sobre agricultura y tildarlo de matemático puro, sin contacto con la ciencia experimental (una acusación a la que Neyman respondió, por descontado, con poca delicadeza). En su momento, Fisher escribió que si la intolerancia a nuevas

ideas era un signo de senilidad, Karl Pearson la había desarrollado desde muy joven, y bien podría decirse que Fisher hizo lo propio, convirtiéndose demasiado pronto en un egocéntrico dinosaurio de la estadística. Siempre se mostró muy poco generoso con Neyman, a pesar de que este lo admiraba y de que su teoría de los intervalos de confianza y del contraste de hipótesis clarificó tanto la mística probabilidad fiducial como las pedestres pruebas de significación. Los roces entre Fisher y Neyman fueron constantes mientras duró su convivencia bajo el techo común del University College, y ni siquiera se calmaron cuando, en 1938, Neyman partió hacia Berkeley, en Estados Unidos. La animadversión entre ambos estadísticos significó la mayor grieta abierta entre los partidarios de la inferencia frecuentista.

> «Fisher a veces publicaba insultos que solo un santo podía perdonar.»
>
> —Leonard «Jimmie» Savage (1976).

Aunque históricamente Neyman publicó su teoría de los intervalos de confianza con posterioridad a la teoría de los contrastes de hipótesis, aquellos son previos a estos desde un punto de vista lógico. Sobre 1930 Neyman ya poseía el germen de la idea, probablemente influido por la aproximación fiducial que Fisher desarrollaba paralelamente (aunque soslayó referirlo). De modo que en 1934 sugirió que mucho más interesante que la estimación puntual era obtener un intervalo dentro del cual se tenía cierta confianza de que se encontrase el parámetro que se quería estimar. Un intervalo de confianza consistía en acompañar la estimación puntual con el margen de error que reflejaba la variabilidad de la estimación. Proporcionar la estimación sin indicar su margen de error era de escasa utilidad y podía ser engañoso. Pero, frente a la tradición de ofrecer la estimación puntual y el error probable (lo que determinaba un intervalo con un nivel de confianza del 50%), Neyman barajaba la posibilidad de construir, mediante el concurso de variables «pivotales», intervalos con cualquier nivel de confianza deseado (pongamos por caso, como es habitual, al

95 o 99%). Para cada nivel de confianza determinado se calculaba su margen de error. Naturalmente, con el nivel de confianza aumentaba el margen de error, aunque otra forma de aumentar la confianza era aumentar el tamaño de la muestra.

Por ejemplo, puede preverse que si extraemos muestras de tamaño 16 de una población que se distribuye normalmente con media μ desconocida y desviación típica 4, entonces con probabilidad 0,95 la media muestral $\bar{X}$ no distará de la media poblacional μ desconocida más de 1,96 unidades. En consecuencia, si al tomar una muestra observamos que $\bar{X} = 40$, puede esperarse que μ se encuentre previsiblemente en el intervalo $40 \pm 1{,}96$ (con un 95% de confianza).

Ahora bien, ¿qué significa la coletilla «al 95% de confianza»? Quiere decir que la estimación por intervalo se ha realizado con un procedimiento que se sabe que a la larga acierta el 95% de las veces. Es como si el intervalo nos lo comunicara una persona que dice la verdad el 95% de las veces; podemos estar bastante seguros, pero no totalmente seguros. Conviene advertir, según insistió Neyman, que si I es un intervalo de confianza concreto al 95%, no se puede decir que la probabilidad de que I contenga el verdadero valor del parámetro θ es 0,95 porque el parámetro θ estará o no estará en I, pero no tiene más opciones, ya que es una constante de valor definido aunque desconocido. Dicho de otra manera, la probabilidad de que I incluya a θ solo puede asumir dos valores: 1 o 0, dependiendo de si θ está o no en I. Sucede que la fórmula que ha permitido construir el intervalo I al sustituir los datos observados posee una probabilidad de 0,95, lo que se interpreta, desde la definición objetiva o frecuencial de la probabilidad, como que el 95% de las muestras producen un intervalo que en efecto contiene el parámetro. Sin embargo, es imposible conocer si nuestro intervalo concreto I es uno de ellos, pero se espera que así sea con un 95% de confianza.

Cuando en 1955 Fisher y Neyman volvieron a cruzar espadas con motivo del artículo incendiario que el primero comunicó a la Real Sociedad de Estadística, Fisher dejó entrever que la concepción de Neyman ponía en peligro su método fiducial, aparte de ser supuestamente una copia degenerada (y ello a pesar de que

los intervalos fiduciales dejan de coincidir con sus hermanos, los intervalos de confianza, cuando se aplican a problemas multiparamétricos como el de Behrens-Fisher). Recordemos que mediante un extraño argumento, Fisher cambiaba el estatus del parámetro θ para hacerlo susceptible de recibir una distribución de probabilidad. Pasaba de suponerlo una constante a una variable aleatoria, una asunción que lo sacaba del paradigma de la estadística clásica y lo sumergía en el marco de la estadística bayesiana. Porque para los bayesianos es posible entender un intervalo de confianza *I* al 95% como que el parámetro θ se encuentra ahí con una probabilidad (subjetiva, credencial) de 0,95.

«De un saco de judías blancas y negras saco un puñado y cuento el número de judías blancas y el número de judías negras y entonces presumo que las blancas y las negras están aproximadamente en la misma proporción en todo el costal.»

— Charles Sanders Peirce sobre el muestreo como base de la inducción.

Mientras que los estadísticos bayesianos contestan a la pregunta de por qué empleamos este intervalo *I* en particular, los estadísticos frecuentistas responden a la pregunta de por qué empleamos intervalos de confianza en general, esgrimiendo que el método de Neyman es un razonamiento deductivo que arroja un 95% de éxitos a largo plazo. La confianza no es una medida de precisión final (atribuible al intervalo numérico construido) sino inicial.

Los contrastes de hipótesis guardan, como en seguida veremos, un nexo fundamental con los intervalos de confianza. Buscando fortalecer las bases lógicas de los test de significación de Fisher, Pearson y Neyman idearon varias mejoras. El *leitmotiv* de su investigación no era otro que el siguiente interrogante: ¿qué hacer si se obtiene un resultado significativo en un test estadístico? De acuerdo, se rechaza la hipótesis nula, pero ¿qué otra hipótesis puede abrazarse? En este sentido las pruebas de significación eran peores que inútiles. No daban ninguna pista.

La teoría de Neyman-Pearson planteaba una elección real entre dos hipótesis rivales. El contraste de hipótesis es un algo-

ritmo para decidir entre dos afirmaciones sobre un parámetro a partir de la información contenida en la muestra. Una será rechazada; la otra, aceptada. Tras formular la hipótesis nula H_0, se formula la hipótesis alternativa H_1, que difiere de la hipótesis de partida. A continuación, se elige el tamaño del test o nivel de significación α deseado, que marca la barrera que juzga qué discrepancias son «demasiado» grandes. Usualmente, suele tomarse $\alpha = 0{,}05$ (el valor complementario al consabido 0,95). Este número determina el riesgo aceptado, esto es, el porcentaje de muestras que tomaremos como significativas para decir que la muestra no es compatible con la hipótesis nula (en este caso, el 5%). Asimismo, se elige el estadístico T del contraste, cuya distribución en el muestreo ha de ser conocida, y que funciona como una medida de la discrepancia entre la hipótesis nula, la hipótesis alternativa y los datos muestrales. Con α y con T se construyen la «región crítica» o «región de rechazo» y la complementaria «región de aceptación de la hipótesis nula» (esta última viene dada por un intervalo de confianza de nivel $1-\alpha$). El hecho de que el valor $T(X)$ observado en la muestra del estadístico del contraste caiga dentro de una u otra dictamina si la diferencia observada es o no significativa, si hay que rechazar la hipótesis nula y aceptar la hipótesis alternativa.

Todo contraste de hipótesis conduce, pues, a aceptar o rechazar la hipótesis nula planteada (aceptando, en este último caso, la hipótesis alternativa). Ahora bien, pueden ocurrir las siguientes situaciones (que aparecen esquematizadas en la tabla):

a) Se acepta la hipótesis nula siendo verdadera. Esta es una decisión correcta.

b) Se rechaza la hipótesis nula siendo falsa. Esta es otra decisión correcta.

c) Se rechaza la hipótesis nula siendo verdadera. Está claro que cometemos un error, que se llama *error de tipo I*. La probabilidad de cometer este error viene dada por el nivel de significación α, fijado de antemano.

d) Se acepta la hipótesis nula siendo falsa. También cometemos un error, que se llama *error de tipo II*. La probabilidad de cometer este error se representa por β, y la probabilidad 1 – β se llama *potencia del contraste*, ya que cuantifica la probabilidad de rechazar la hipótesis nula cuando es falsa.

	Naturaleza de la hipótesis nula H_0	
Decisión	Verdadera	Falsa
Rechazar H_0	*Error de tipo I*	Correcta
No rechazar H_0	Correcta	*Error de tipo II*

Neyman y Pearson demostraron que en bastantes circunstancias, una vez fijada la probabilidad α de error de tipo I (esto es, asumiendo la interpretación frecuentista del muestreo repetido, una vez acotado el porcentaje de veces que tomaremos una decisión equivocada, al rechazar la hipótesis nula cuando es verdadera), es posible construir y utilizar contrastes de máxima potencia, es decir, contrastes que minimizan la probabilidad β de error de tipo II al tiempo que maximizan la potencia del test, su sensibilidad o capacidad para detectar que la hipótesis nula es falsa. En un célebre lema publicado en 1933, Neyman y Pearson probaron que en el caso de hipótesis rivales simples (que asignan valores específicos al parámetro desconocido) existe automáticamente una clase de test óptimos, de bajo tamaño y máxima potencia: los basados en la razón de verosimilitudes (ver anexo al final del libro). Según dejaron escrito en 1933:

> Sin esperar conocer si cada hipótesis por separado es verdadera o falsa, buscamos reglas que gobiernen nuestro comportamiento con respecto a ellas, de modo que a la larga no estemos frecuentemente equivocados.

De acuerdo con el planteamiento de Neyman y Pearson, un contraste de hipótesis no es más que una regla de decisión. Si uno se comporta conforme al procedimiento diseñado, a la larga rechazará la hipótesis nula cuando sea verdadera no más, digamos, que cinco veces de cada cien y, además, dispondrá de evidencia de que la rechazará con la suficiente frecuencia cuando sea falsa. Los test estadísticos no son, por tanto, reglas de inferencia inductiva, sino de comportamiento inductivo. Su propósito no es fundamentar nuestras creencias, sino ajustar nuestra conducta a los datos observados. No es posible averiguar si la hipótesis nula es verdadera o falsa. Pero, en cambio, sí es factible comportarnos respecto a ella de manera que a largo plazo no erremos con demasiada frecuencia. Frente a Fisher, Neyman y Pearson sostenían que lo que es inductivo no es el razonamiento sino la acción. El objeto de la estadística era emplear la experiencia como guía para actuar apropiadamente. Ni más, ni menos.

Los procesos de control de calidad en la producción industrial siguieron de cerca esta visión. Así, durante la Segunda Guerra Mundial, los contrastes de hipótesis sirvieron para la selección de bastimentos en la Armada estadounidense, ya que inspeccionando una muestra de cada lote podía tenerse la confianza de seleccionar correctamente al menos el 95% de los lotes no defectuosos a largo plazo. Egon Pearson escribió, de hecho, un libro sobre la materia que pereció quemado en uno de los primeros *raids* sobre Londres. Pero fue la emigración de Neyman a Estados Unidos en 1938 lo que facilitó que esta constelación de ideas cruzara el Atlántico y terminara sedimentando en la teoría matemática de la decisión esbozada hacia 1950 por el malogrado Abraham Wald (fallecido tempranamente en un accidente de avión).

En múltiples ocasiones Neyman sostuvo la tantalizante doctrina de que la inferencia inductiva es imposible y debemos contentarnos con la conducta inductiva. Una opinión contundente que le convirtió en el villano de las disputas filosóficas de la estadística. A su entender, la estadística matemática no hacía justicia al presunto carácter inductivo de la empresa científica, ya que su entramado era puramente deductivo. Del mismo modo que los bayesianos y sus epígonos tomaban como premisa una distribu-

ción *a priori* de probabilidad, Fisher partía siempre de la función matemática de verosimilitud o de una distribución en el muestreo deducida con anterioridad. Los intervalos de confianza, por su parte, se obtenían razonando sobre las propiedades de ciertas variables aleatorias. Y los contrastes de hipótesis eran meras reglas de comportamiento, donde no cabía la inferencia, ni inductiva ni deductiva, porque había probabilidades de error. La lógica se resolvía, empero, en decisión.

A juicio de Fisher, tanto Neyman como Pearson habían desvirtuado íntegramente su invención; porque el objetivo de un test de significación —como explicamos en el capítulo 3— no era decidir entre dos hipótesis alternativas, sino comprobar si una observación acreditaba o no la hipótesis nula. Sus queridos test se habían transformado en vulgares recetas de aceptación. Mientras que las pruebas de significación se construían tomando como referencia una única hipótesis y su objeto era validar el modelo estadístico subyacente, los contrastes de hipótesis consideraban dos hipótesis rivales y su propósito principal era decantarse por una de ellas.

Además, para Fisher, Neyman y Pearson habían formalizado las pruebas de significación en un marco (supuestamente) confuso, ya que el resultado de una de estas pruebas venía dado por el p-valor, que medía hasta qué punto los datos no contradecían la hipótesis nula, y no por la decisión de aceptar la hipótesis nula o la hipótesis alternativa. No era lo mismo informar del p-valor, como medida de la evidencia aportada por la muestra, que de la aceptación o el rechazo de la hipótesis nula, con la consiguiente (falsa) creencia de que esta hipótesis era verdadera o falsa simplemente porque no/sí contradecía los datos observados. De hecho, la utilización del p-valor permite que todos los estadísticos a los que se les facilite la misma muestra obtengan idéntico resultado. En cambio, dos estadísticos que informen del resultado de un contraste pueden llegar, a partir de la misma muestra, a resultados distintos si utilizan dos tamaños diferentes, dos α distintos. La razón estriba en que el p-valor es una propiedad de la muestra, mientras que el tamaño α es una propiedad del test.

Al respecto, Fisher protestaba enfadado que la interpretación del nivel de significación α del test como frecuencia de una

Cuestiones candentes en la teoría de Neyman-Pearson



Jerzy Neyman.

A pesar de que los contrastes de hipótesis han sido universalmente aceptados, presentan ciertos déficits técnicos que no deben dejar de señalarse. Primeramente, muchos investigadores creen que para un α fijo, el rechazo de la hipótesis nula, caso de producirse será más evidente conforme mayor sea el tamaño muestral *n*. Sin embargo, esto no es así. Si se quiere contrastar si la producción media de una máquina es de 5000 unidades/día y se toma una muestra grande (una serie larga de observaciones diarias), es bastante probable que se detecte una diferencia estadísticamente significativa y se rechace que la media es 5000. Pero la conclusión bien puede ser que la media es, entonces, de 5000 + 0,00001, una diferencia perfectamente irrelevante en la práctica. Como la región crítica depende del tamaño muestral, el valor por encima del cual se rechaza la hipótesis nula de que la media es 5000 se acerca a 5000 según aumenta *n* (puesto que la media observada ha de estar muy próxima a la media teórica si la muestra es grande). Un efecto pequeño en una muestra grande puede ser tan decisivo como un efecto grande en una muestra pequeña. Para evitar este engorro, hay quienes sugieren ajustar el tamaño del test en función del tamaño de la muestra.

decisión equivocada en muestras repetidas de la misma población pervertía la lógica intrínseca a las pruebas de significación, porque el científico natural generalmente no dispone de muestras repetidas. La analogía que empleaban Neyman y Pearson entre el muestreo repetido y la toma reiterada de decisiones solo funcionaba si se asimilaba el contraste de hipótesis con la aceptación industrial de lotes de muestras. Aún más, la expresión *error de segundo tipo* parecía sugerir la posibilidad de aceptar como verdadera la hipótesis nula por error, cuando la realización de una prueba de significación nunca autorizaba a tomarla como verdadera.

En segundo lugar, como consecuencia del papel privilegiado de la hipótesis nula (ya que α se fija con anterioridad), en ocasiones se tiende a aceptar la hipótesis nula incluso cuando los datos no encajan bien con esta hipótesis. Es más, la obligatoriedad de decidir entre la hipótesis nula y la hipótesis alternativa a veces conduce a tomar decisiones basándose en datos muestrales que encajan igual de mal con ambas hipótesis, algo que con el enfoque bayesiano no pasa (en el anexo al final de libro abundamos en esta cuestión).

La potencia del test

Neyman enfatizaba que la no significatividad de un test para rechazar la hipótesis nula no lleva necesariamente a verla confirmada, ya que esto depende de la potencia del test, de que sea lo suficientemente alta. Algunos estadísticos apuntan que la fuerza con que la hipótesis nula se ve confirmada por la muestra puede evaluarse mediante una cantidad que denominan *severidad*, y que jugaría un papel análogo al p-valor. Mientras que el p-valor se definía —como vimos en el tercer capítulo— por la probabilidad $P(T \geq T(X)|H_0)$, la severidad se definiría por $P(T \geq T(X)|H_1)$. Cuanta más alta fuese esta probabilidad, más «duro» o «severo» habría sido el test en el sentido de ser capaz de discernir si la hipótesis nula era falsa. Un experimento confirmaría una hipótesis si y solo si suponía un intento serio por refutarla. Por último, en tercer lugar, cuando las hipótesis no son simples sino compuestas, el lema fundamental no se verifica y la búsqueda del test uniformemente más potente no siempre existe, con lo que no es fácil controlar simultáneamente las dos probabilidades de error. Ya en su momento Fisher puso de relieve que, para rizar el rizo, el cálculo del error del segundo tipo y, por tanto, de la potencia del contraste, no siempre es accesible, dado que la hipótesis alternativa puede no estar unívocamente determinada.

Las diferencias entre ambas teorías no eran tanto matemáticas, numéricas, como lógicas y filosóficas. En el polémico artículo presentado por Fisher en 1955 a la Real Sociedad de Estadística, el estadístico británico atacó furibundamente a Neyman por dejarse seducir por el «pragmatismo norteamericano», por mostrarse más preocupado por acelerar la producción que por extraer conclusiones estadísticas correctas. El matemático polaco había malinterpretado la inferencia estadística al constreñirla, como decía literalmente Fisher, al ámbito de los esclavos de Wall Street y del Kremlin, pero no de los científicos libres en pos de la verdad. Neyman había cortado el nudo gordiano de la lógica de

la inferencia inductiva de la que hablaba Fisher al calificarla como ilusoria. Pero en su ceguera había confundido el control de calidad con la inferencia científica, al científico con el comerciante. El «comportamiento inductivo» le parecía a Fisher una evasión para no afrontar el problema realmente existente del «razonamiento inductivo». Fisher no quería hacer dinero sino aprender del experimento.

La réplica que Neyman no tardó en escribir comenzaba salvando al desgraciado Wald de las invectivas de Fisher: la relación de la inferencia estadística con la teoría de la decisión pergeñada por Wald era la de la táctica con la estrategia. A continuación, Neyman defendía su enfoque mediante hipótesis alternativas, llegando a subrayar que el célebre test de la catadora de té estaba mal diseñado si no se indicaba contra qué se quería probar la hipótesis nula (es decir, si no se precisaba numéricamente la habilidad de la dama, suponiendo que la tuviera, en la hipótesis alternativa). En lo tocante al tema central de discusión, Neyman se reafirmaba en que el comportamiento inductivo solventaba de una vez por todas el problema irresoluble de la inferencia inductiva.

Con el tiempo, el matemático polaco llegó a referirse a la conducta inductiva —incluso en presencia del filósofo Carnap— como un concepto mayor de la filosofía de la ciencia actual, hallando sus raíces en Gauss y Laplace. En cierto modo las voces de Neyman y Popper se confunden en este punto al afirmar ambos que no existe método inductivo de razonamiento alguno. Si para Popper los posibles resultados de una prueba experimental son la falsación o, en su defecto, la corroboración de la teoría científica, para Neyman lo son el rechazo o la aceptación de la hipótesis nula (aunque como en el caso de Fisher, Popper apenas citó a Neyman).

Por alusiones, Egon Pearson también hubo de terciar en la polémica, aunque a diferencia de Neyman se resistió a bajar a la arena filosófica, limitándose a aducir que la jerga de la toma de decisiones pertenecía más a Neyman que a sí mismo. La buena sintonía entre ambos matemáticos se había prácticamente terminado cuando el segundo partió rumbo a Estados Unidos.

USOS Y ABUSOS DE LOS MÉTODOS ESTADÍSTICOS

El sincretismo metodológico reinante es responsable de bastantes errores cometidos en el empleo de las herramientas estadísticas. Algunos de los más habituales son los siguientes:

1. En el análisis exploratorio de datos suele usarse la media como medida canónica de centralización, que agrupa las observaciones, cuando la mediana es en general más recomendable por cuanto presenta menor volatilidad, esto es, menor sensibilidad a valores extremos.

2. En el estudio de la regresión habitualmente se toma un coeficiente de correlación lineal de 0,6 como fiable, cuando puede demostrarse que el modelo subyacente solo explica en este caso el 36 % de las observaciones.

3. Una ilusión permanente, fruto del pastiche que ha fraguado en torno a los test estadísticos, es creer que estos se apoyan en el siguiente silogismo: «Si la hipótesis nula es correcta, entonces la muestra *X* no puede observarse. Hemos observado *X*, luego la hipótesis de partida es falsa». Sin embargo, los test descansan sobre un silogismo a lo sumo probable: «Si la hipótesis nula es correcta, entonces la muestra *X* es altamente improbable. *X* ha sido observada, luego la hipótesis es altamente inverosímil».

4. La consagración de la contrastación estadística como modo de tomar decisiones dicotómicas conlleva que a veces, basándose en el criterio del $\alpha = 0{,}05$, se acepte la hipótesis nula para un p-valor de 0,051 y, en cambio, se rechace para 0,049. Asimismo, un resultado estadísticamente significativo al nivel, pongamos, del 0,001 suele interpretarse como que la hipótesis alternativa ha recibido un apoyo del 0,999; pero que no haya evidencia en contra suya no quiere decir que la tenga a favor.

5. Otro error muy extendido es confundir el p-valor, es decir, la probabilidad de observar la muestra extraída suponiendo que la hipótesis nula es verdadera, con la probabilidad de que la hipótesis nula sea verdadera a la vista de la muestra observada (una probabilidad solo calculable mediante el teorema de Bayes). Esta inversión ilegal de los términos es lo que se conoce como *falacia del fiscal*: si eres culpable, es lógico que todas las pruebas apunten a ti; pero que todas las pruebas apunten a ti, no quiere decir que *ipso facto* seas culpable, como suelen inferir erróneamente los fiscales.

6. Finalmente, hay que anotar que la potencia del contraste es la gran olvidada de la teoría. Entre los investigadores ha fructificado la creencia de que si un test no resulta significativo, entonces la hipótesis nula ha sido corroborada; pero esto no puede afirmarse a la ligera sin antes calcular la función de potencia del test, que mide su capacidad para detectar discrepancias.

La disputa entre Fisher y Neyman en 1955 inauguró toda una serie de controversias en la que ya no intervendrían solo estadísticos, sino también filósofos interesados por la inferencia científica, que subrayarían que la teoría de los contrastes de hipótesis es idónea para poner a prueba una hipótesis pero no para evaluar el respaldo que recibe esta hipótesis una vez realizado el experimento. En otras palabras, la inferencia clásica es la más adecuada para someter una hipótesis al dictado de la experiencia; pero, una vez que la naturaleza habla, la inferencia bayesiana ha de recoger el testigo (ya que posibilita la comparación entre las alternativas por medio de sus probabilidades *a posteriori*).

Ahora bien, el propósito principal de los contrastes de hipótesis no es medir el grado de apoyo que recibe una hipótesis a partir de la muestra observada, sino evaluar la discrepancia de esta hipótesis con los datos. En el esquema clásico, las probabilidades entran en juego como probabilidades de error, no como probabilidades de hipótesis. Al igual que el nivel de confianza, las probabilidades de error funcionan como medidas de precisión inicial, no final. Los test ideados por Fisher, Neyman y Pearson no pueden transformarse en lo que no son. No se les puede pedir lo que no pueden dar.

Y, sin embargo, a día de hoy, ha triunfado el más vivo eclecticismo metodológico, en especial en el campo de las ciencias sociales, donde las pruebas de significación de Fisher y los contrastes de hipótesis de Neyman-Pearson, e incluso en ocasiones los modelos bayesianos, cohabitan en una amalgama viable a escala técnica pero irreconciliable a escala conceptual. A partir de los años sesenta del pasado siglo las teorías de Fisher y de Neyman-Pearson comenzaron silenciosamente a conformar un oscuro híbrido cuyo uso se ha trivializado, convirtiéndose en un ritual mecánico. Bajo el pensamiento débil de que ¡todo vale! («cualquier método estadístico es un instrumento válido», «no hay que entrar en disquisiciones lógicas»), se oculta un problema de calado filosófico con repercusiones a la hora de plasmar e interpretar los resultados, porque no es lo mismo informar del p-valor que de la distribución *a posteriori* o del tamaño del test, la potencia del contraste y la decisión tomada.

FUMAR PERJUDICA GRAVEMENTE LA SALUD

Hacia 1920 se observó un gran incremento de los fallecimientos por cáncer de pulmón. Aunque existían trabajos previos sobre la posible relación entre este tipo de cáncer y el hábito de fumar, en la década de 1950, gracias a los trabajos de Richard Doll (1912-1905) y Austin Bradford Hill (1897-1991), la cuestión cobró un verdadero interés y propició agrios debates en la opinión pública. Estos epidemiólogos fueron los artífices de la extensión de los principios fisherianos del diseño de experimentos a la investigación clínica.

Doll y Hill publicaron un estudio estadístico donde los casos los constituían los pacientes que ingresaban en ciertos hospitales con diagnóstico de cáncer de pulmón, mientras que el «grupo control» estaba formado por pacientes cuyo ingreso se debía a otras causas. Mediante el análisis de las historias clínicas de los enfermos que ya tenían o que desarrollaron este cáncer, estimaron que la incidencia del mismo en los fumadores era entre 11 y 20 veces mayor que en los no fumadores. Su conclusión era, *de facto*, estadísticamente significativa al nivel del 0,001.

Sin embargo, estos trabajos recibieron numerosas objeciones de personalidades tan respetadas como Jerzy Neyman. Pero quizá el principal paladín de las críticas fue nada menos que Fisher (a quien distinguimos en muchas fotografías pipa en mano). Este inveterado fumador, que incluso sirvió como consultor de alguna compañía tabacalera, publicó varios artículos y un panfleto cuestionando la relación entre cáncer, cigarrillos y estadística.

Una de las pegas que Fisher esgrimió fue que el estudio demostraba que los fumadores presentaban un mayor riesgo de padecer cáncer de pulmón, pero esto no implicaba que la causa fuese necesariamente el tabaco. Que A y B estén directamente correlacionadas no quiere decir que A sea la causa de B, pues bien podría ser que B fuera la causa de A (que el cáncer de pulmón motivara el hábito de fumar) o que existiese un factor C que fuese la causa común de A y B (que las personas que adquieren el hábito de fumar tuviesen algo en la estructura genética que las hiciera

propensas a caer en la adicción al tabaco y, a la vez, contraer un cáncer; una posibilidad que Fisher barajaba amparándose en datos extraídos de gemelos). El estadístico inglés comparaba la correlación descubierta por Doll y Hill con la correlación engañosa que mediaba entre la evolución de la tasa de divorcios y la importación de manzanas.

Fisher añadía que, a diferencia de los experimentos agrónomos o los estudios sobre vacunas, el estudio de Doll y Hill no se ajustaba al diseño experimental, sino que era un mero estudio prospectivo, porque la división en dos grupos —casos y controles— no se había producido aleatoriamente, sino que venía dada y, por tanto, sujeta a factores externos difíciles de bloquear. Es más, subrayaba que si uno separaba a los fumadores en dos grupos, los que inhalan el humo y los que no, los que no inhalaban el humo eran curiosamente los que más padecían cáncer de pulmón. Fisher escenificaba la conclusión real del estudio con el siguiente consejo: «fumar perjudica la salud, pero si tienes que fumar, mejor traga el humo».

Los años sucesivos conocieron una multiplicación de estudios prospectivos, así como de experimentos con animales que corroboraron fuera de toda duda la tesis de Doll y Hill (y mostraron que, pese a lo que por error arrojaba el primer estudio, inhalar el humo resulta fatal). A medida que la evidencia se fue acumulando Neyman cambió de opinión, pero Fisher permaneció irreductible en su posición.

LA ESTADÍSTICA EN EL SIGLO XXI

Ronald Aylmer Fisher nunca ocupó una plaza como estadístico en la universidad. En 1957 tomó la decisión de abandonar la cátedra de Genética en la Universidad de Cambridge y, dos años después, se incorporó como investigador emérito a un complejo científico e industrial ligado a la Universidad de Adelaida (Australia). Este genio de temperamento, que había sido nombrado sir por la reina Isabel II en el año 1952, encontró la muerte el 29



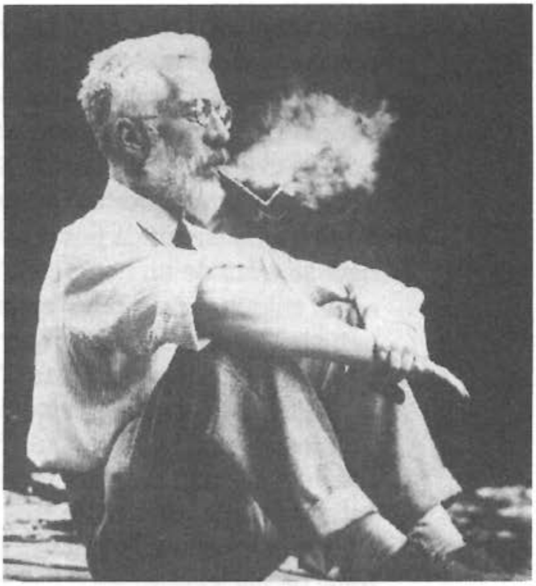




FOTO SUPERIOR IZQUIERDA:
Fisher durante una visita a la India, acompañado por C.R. Rao, un estadístico indio que se doctoró en Cambridge bajo la tutela de Fisher.

FOTO SUPERIOR DERECHA:
Una de las polémicas que sostuvo Fisher, fumador empedernido, fue no reconocer la vinculación entre el hábito de fumar y el cáncer de pulmón.

FOTO INFERIOR:
En 2002 se inauguró la «placa azul» dedicada a R.A. Fisher, en presencia de tres de sus hijos, June, Margaret y Harry.

de julio de 1962, a los setenta y dos años, como consecuencia de un cáncer de colon.

Los avances que Fisher impulsó le otorgan un puesto de honor en el panteón de los estadísticos. Gracias a él, la estadística es la matriz de muchas ciencias experimentales. En tanto que la experimentación produce datos varios, precisa de la estadística. Todo hecho científico posee un carácter ineludiblemente estadístico: se trata de un compendio de observaciones repetidas, que están sujetas a factores y errores de naturaleza aleatoria. La estadística interviene en la descripción, modelización, explicación y predicción de estos datos. Y lo hace, en general, cumpliendo las siguientes etapas: planteamiento de un modelo adecuado al problema utilizando el cálculo de probabilidades; diseño del experimento; descripción y análisis de los datos muestrales recogidos; estimación de los parámetros desconocidos del modelo poblacional; contraste de hipótesis sobre el modelo; reajuste de este y toma de decisiones.

«Lo mejor de ser estadístico es que puedes meterte en cualquier jardín.»

— John W. Tuckey.

Al igual que otros estilos de razonamiento científico (el geométrico de las ciencias matemáticas, el hipotético-deductivo de las ciencias físicas, el experimental de las ciencias de laboratorio, el taxonómico de las ciencias naturales y el histórico-genético de las ciencias humanas), hay un estilo propio de operar, pensar y actuar enlazado a la ciencia estadística, que se caracteriza por una fértil dialéctica entre razonamiento y experimentación.

La aplicación de los métodos estadísticos se ha extendido a áreas tan diversas como la ingeniería, la economía, la medicina o la psicología. En la actualidad, tanto los filtros de *spam* de nuestro ordenador como la observación de cúmulos estelares, la detección del fraude fiscal o el análisis de las causas de accidentes como el del *Challenger* en 1986 emplean técnicas estadísticas.

La difusión de la estadística, de la que Ronald Aylmer Fisher fue partícipe privilegiado, no solo ha provocado que el mapa se pliegue mejor al territorio, sino también que a resultas de ello el territorio —nuestro mundo globalizado— se haya visto transformado hasta límites insospechados por culpa de la introducción del mapa. Habitamos un mundo estadístico en el que el mapa se confunde con la realidad.

Anexo

TESTANDO A FISHER, NEYMAN Y BAYES

El objetivo de este anexo es presentar matemáticamente cómo cada una de las tres escuelas estadísticas posee un enfoque muy distinto a la hora de analizar un mismo caso de estudio. Por medio de un ejemplo numérico sencillo, el lector podrá comprobar cómo cada una de estas filosofías de la estadística interpreta los cálculos probabilísticos de una manera sutilmente diferente.

Supongamos que un parámetro poblacional θ desconocido solo puede tomar dos valores: 0 o 1. Supongamos, además, que los datos muestrales X que observaremos únicamente tienen cuatro posibles resultados: 1, 2, 3 o 4. La siguiente tabla recoge las probabilidades $P(X|\theta)$ de observar cada resultado muestral en función de los valores del parámetro:

$P(X\mid\theta)$	$X=1$	$X=2$	$X=3$	$X=4$
$\theta=0$	0,980	0,010	0,005	0,005
$\theta=1$	0,098	0,900	0,001	0,001

TEST DE SIGNIFICACIÓN DE FISHER

Queremos poner a prueba la hipótesis nula de que $\theta=0$. De acuerdo con Fisher, no hacemos referencia a hipótesis alterna-

tiva alguna ($\theta = 1$), ya que nuestro objetivo no es decidir entre dos hipótesis rivales, sino validar el modelo estadístico subyacente que presupone ese valor para el parámetro desconocido. Si recordamos del capítulo 3, el p-valor se definía como la probabilidad $P(T \geq T(X)|H_0)$, lo que en este caso discreto se adapta como la probabilidad de observar un valor igual o más raro que el valor efectivamente observado bajo la hipótesis de que $\theta = 0$. Con esto en mente, ¿qué inferiremos si observamos que $X = 2$?

Por lógica, mirando la tabla anterior, como la probabilidad de observar este resultado muestral suponiendo que $\theta = 0$ es muy baja (de solo 0,010), el p-valor ha de ser pequeño. En efecto, vale $0{,}010 + 0{,}005 + 0{,}005 = 0{,}02$, que al ser menor que el consabido límite de 0,05, apunta a que la hipótesis nula no encaja con el dato observado y, por tanto, ha de ser rechazada.

¿Y si observamos $X = 3$? Entonces el p-valor vale $0{,}005 + 0{,}005 = 0{,}01$, lo que conduce a rechazar la hipótesis nula de que $\theta = 0$ con mayor significación. Finalmente, si se observa $X = 1$ (el dato para el que la hipótesis nula encaja muy bien, ya que este dato se observa con probabilidad 0,980), el p-valor es $0{,}980 + 0{,}010 + 0{,}005 + 0{,}005 = 1$, lo que de ningún modo contradice la hipótesis nula. En resumen, el p-valor es la medida matemática que informa en los test de significación de hasta qué punto la muestra refuta la hipótesis de partida. Pero nada dice de en qué grado permite inferirla o confirmarla.

CONTRASTE DE HIPÓTESIS DE NEYMAN-PEARSON

Consideramos la hipótesis nula H_0: $\theta = 0$ versus la hipótesis alternativa H_1: $\theta = 1$. El propósito del contraste es decidir entre ambas. Intuitivamente, consultando la tabla, si observamos $X = 1$, aceptaremos la hipótesis nula. En cambio, si observamos $X = 2$, nos inclinaremos por rechazarla, aceptando la hipótesis alternativa. Cuando $X = 3$ o 4, la decisión no está tan clara.

Como explicamos en el capítulo 5, la teoría de Neyman-Pearson comienza balanceando las dos probabilidades de error. En primer lugar, se fija el tamaño o nivel dé significación α del test,

que acota la probabilidad del error de tipo I (esto es, la frecuencia con que tomamos la decisión equivocada de rechazar la hipótesis nula cuando es verdadera). A continuación, se busca aquel test con menor probabilidad de error de tipo II (de aceptar la hipótesis nula cuando es falsa) o, equivalentemente, con mayor potencia, es decir, con mayor probabilidad de rechazar la hipótesis nula cuando es, en efecto, falsa. Según demostraron Neyman y Pearson en un famoso lema, los test óptimos (tamaño pequeño, máxima potencia) se basan en la razón de verosimilitudes, es decir, en el cociente $P(X|\theta=1)/P(X|\theta=0)$, que se obtiene dividiendo las probabilidades (verosimilitudes) de la tabla:

	$X=1$	$X=2$	$X=3$	$X=4$
$\frac{P(X\mid\theta=1)}{P(X\mid\theta=0)}$	0,1	90	0,2	0,2

Es fácil ver que la razón de verosimilitudes va a conducir al rechazo de la hipótesis nula y la aceptación de la hipótesis alternativa cuando $X=2$ (como era de esperar), ya que el cociente toma un valor muy grande (la verosimilitud de la hipótesis alternativa es 90 veces la de la hipótesis nula). Cuando $X=1$, mantendremos la hipótesis nula, porque el cociente toma el valor más pequeño (0,1). Y si $X=3$ o 4, la decisión dependerá del tamaño α elegido del test, puesto que los resultados muestrales encajan prácticamente igual de mal con ambas hipótesis (la probabilidad de observar 3 o 4 era baja con ambas hipótesis). Así, puede demostrarse que con $\alpha=0{,}01$ la región crítica para H_0: $\theta=0$ solo contiene a $X=2$. En consecuencia, para $X=3$ o 4 retenemos la hipótesis nula. La potencia de este test vendría dada por la probabilidad $P(X=2|\theta=1)$ de rechazar la hipótesis nula cuando la hipótesis alternativa es verdadera, que arroja un valor (consultando la tabla inicial) de 0,900. Por consiguiente, este test muestra una gran potencia, en otras palabras, una gran capacidad para detectar cuándo la hipótesis nula es falsa. En concreto, si se observa $X=1$ (un resultado no significativo), la «severidad» del test viene dada por

$P(T \geq T(X)|\theta = 1) = 0{,}900 + 0{,}001 + 0{,}001 + 0{,}098 = 1$, lo que ofrece una evidencia excelente para inferir la hipótesis nula frente a la alternativa.

Sin embargo, con $\alpha = 0{,}02$, la región crítica incluye a $X = 2$, 3 y 4, por lo que rechazaríamos la hipótesis de partida en todas estas circunstancias, a pesar de que la hipótesis nula es más verosímil que la hipótesis alternativa cuando $X = 3$ o 4. Como se ha dicho, los datos muestrales 3 y 4 constituyen sucesos raros bajo cualquiera de las dos hipótesis rivales, pero la obligatoriedad de decidir entre una y otra fuerza siempre a tomar una decisión en la teoría de Neyman-Pearson. Esta es una de las críticas que los partidarios de la inferencia bayesiana suelen hacer a los defensores de la inferencia frecuentista, ya que con el enfoque bayesiano, como enseguida comprobaremos, esto no siempre pasa.

No obstante, una línea de defensa de los estadísticos clásicos es la apelación a la noción de *severidad*. De este modo, por ejemplo, la decisión de aceptar la hipótesis alternativa cuando $X = 3$ (un resultado significativo) no es un indicio que permita inferir esta hipótesis fuera de toda duda razonable, ya que la severidad del test para con H_1 es —aunque la justificación de la fórmula excede el alcance del libro— $P(T \leq T(X)|\theta = 1) = 0{,}098 + 0{,}001 + 0{,}001 = 0{,}1$ (muy pequeña). La severidad del test es muy baja porque la potencia es muy alta, exactamente de 0,902. Tomemos un ejemplo ilustrativo para explicar por qué se da esta relación: si usamos una red muy tupida para pescar, tendremos muchas oportunidades de pescar un pez y, en consecuencia, de rechazar la hipótesis nula de que el lago no contiene peces (alta potencia); pero si logramos pescar, como los agujeros de la red son tan pequeños y capturan casi todo, no podremos saber si el pez es pequeño o grande y, por tanto, confirmar una hipótesis alternativa con respecto al tamaño de los peces del lago (baja severidad). En suma, la observación del dato muestral 3 conduce a rechazar H_0 (ya que para $\theta = 0$ es muy improbable observarlo), pero de aquí no se desprende necesariamente la verdad de H_1 (de que $\theta = 1$, porque para este valor también es muy improbable observarlo). El lector perspicaz puede estar preguntándose por qué no consideramos el típico $\alpha = 0{,}05$. La razón es que requeriría, al tratarse de un ejemplo discreto, la

introducción de un «test aleatorio», lo que complicaría en exceso la discusión.

INFERENCIA BAYESIANA

El análisis bayesiano precisa de postular una distribución *a priori* sobre θ. A continuación, mediante la aplicación del teorema de Bayes (que presentamos en el capítulo 1), pueden combinarse estas probabilidades *a priori* con las verosimilitudes a fin de obtener las probabilidades *a posteriori* que permitan decantarnos entre H_0 y H_1. Vamos a considerar dos priores distintas. La primera será uniforme, es decir, neutral, no informativa, otorgando la misma probabilidad a los dos posibles valores de θ: $P(\theta=0)=P(\theta=1)=1/2$. La segunda, en cambio, otorgará cinco veces más credibilidad al valor $\theta=1$: $P(\theta=0)=1/6$; $P(\theta=1)=5/6$. Así pues, para cada uno de los dos posibles valores de θ, la probabilidad *a posteriori* vendrá dada por la fórmula de Bayes, expresada a continuación:

$$P(\theta|X)=\frac{P(\theta)\cdot P(X|\theta)}{P(0)\cdot P(X|0)+P(1)\cdot P(X|1)}$$

Según puede calcularse, en el primer caso, si tomamos la distribución uniforme y observamos $X=1$, la probabilidad *a posteriori* es claramente favorable a la hipótesis nula frente a la alternativa: $P(\theta=0|X=1)=0{,}91$, mientras que $P(\theta=1|X=1)=0{,}09$. Si observamos $X=2$, la probabilidad *a posteriori* favorece, como se esperaba, la hipótesis alternativa: $P(\theta=0|X=1)=0{,}01$ frente $P(\theta=0|X=1)=0{,}99$. Pero, ¿qué sucede si $X=3$ o 4 (los valores muestrales que planteaban problemas a la teoría clásica)? Tomando $X=3$, se comprueba que la regla de Bayes se inclina por la hipótesis nula frente a la alternativa: $P(\theta=0|X=3)=0{,}83$ y $P(\theta=1|X=3)=0{,}17$. Sin embargo, cuando introducimos la segunda prior (que otorga más peso *a priori* a $\theta=1$ que a $\theta=0$), el panorama cambia radicalmente: $P(\theta=0|X=3)=0{,}50$ y $P(\theta=1|X=3)=0{,}50$. ¡En equilibrio! Como puede observarse, la

elección de la prior resulta decisiva en el enfoque bayesiano y decanta la balanza hacia uno u otro lado.

INFERENCIA CLÁSICA

Por último, nos gustaría mostrar con otro ejemplo cómo opera la inferencia clásica en la vida real. Vamos a inspirarnos en una aplicación que Fisher extrajo del célebre artículo de Student de 1908. Se desea testar el poder de un nuevo medicamento para inducir al sueño, y se ha medido el número de horas de descanso que 10 pacientes han ganado o perdido con esta droga hipnótica con respecto a no usarla. Es lo que se llama una *muestra con observaciones pareadas*, porque las comparaciones se realizan sobre las mismas 10 personas (si se tratase de 10 personas distintas en cada caso, se trataría de dos *muestras independientes*, que requieren de otro test estadístico algo más complejo; con muestras apareadas pueden captarse efectos invisibles para las muestras independientes). Estas han sido las diferencias observadas con el uso: +1,2; +2,4; +1,3; +1,3; +0; +1; +1,8; +0,8; +4,6; +1,4. A simple vista, parece que el sedante es efectivo, pero podría ser que el efecto se debiese al azar y no a la dosis. La media muestral $\bar{X}$ vale +1,58 (lo que refuerza nuestra opinión), pero nos gustaría contrastar la hipótesis nula de que la media poblacional μ es 0 frente a la hipótesis alternativa $\mu \neq 0$. En otras palabras, la hipótesis de que si el medicamento se suministrase a toda la población no se detectaría efecto alguno versus la hipótesis de que sí lo hay.

Supongamos que el número de horas de sueño que se ganan o se pierden con el sedante sigue una distribución normal de media μ y desviación típica σ desconocidas. A partir de los datos de la muestra, queremos precisamente estimar el efecto medio μ del medicamento sobre toda la población. Se sabe por el teorema central del límite que para muestras grandes ($n > 30$), en condiciones muy generales,

$$\frac{\text{estimador-parámetro}}{\text{desviación típica del estimador}} \sim \text{distribución normal estándar.}$$

Para el caso de la estimación de la media poblacional μ con muestras pequeñas en poblaciones normales, si conociéramos la desviación típica poblacional σ, aún podríamos emplear la aproximación normal. Con una confianza del 95%, la media poblacional μ se encontraría de la media muestral $\bar{X}$ a menos de 1,96 veces la desviación típica poblacional σ dividida por la raíz cuadrada del tamaño muestral n. O como gustaba decir a Fisher, solo una vez de cada veinte excedería estos límites, fijados para el nivel clásico de significación del 5%.

Cuando no se conocía σ (lo más frecuente), el astrónomo F.W. Bessell conjeturó que podía sustituirse su conocimiento por el de la desviación típica muestral corregida $\hat{S}$ (la raíz cuadrada de la cuasivarianza muestral, definida en el capítulo 3, y que en nuestro ejemplo vale 1,23) y sucumbió a la tentación de decir que los valores aceptables eran aquellos que no excedían de:

$$\pm 1{,}96 \cdot \frac{\hat{S}}{\sqrt{n}}.$$

Sin embargo, esta estimación, que hizo fortuna durante el siglo XIX, obviaba el hecho de que $\hat{S}$ está sujeta a las variaciones azarosas del muestreo, por lo que en unas ocasiones será mayor y en otras menor que σ. Student fue el primero en percibir que este olvido afectaba a las conclusiones con muestras pequeñas, reparando en que la distribución normal (de donde procede el $\pm 1{,}96$) no podía emplearse. En su lugar había que usar una nueva distribución, la t de Student, cuyas colas de valores extremos decrecen mucho más lentamente. En consecuencia, el refinamiento de la inferencia pasaba por usar como valor adecuado $\pm 2{,}262$ (al 5% de significación). Curiosamente, Student envió las tablas de su distribución a Fisher con el comentario: «Probablemente sea la única persona que las use jamás». El paso del tiempo ha demostrado, contra la opinión de Karl Pearson, la ubicuidad de la t de Student, ya que su uso es generalmente válido con independencia de que la distribución de partida sea normal.

Resumiendo, si desconocemos σ, hay que emplear la aproximación que descubrió Student, a la que tanto juego sacó Fisher:

$$\frac{\text{media muestral} - \text{parámetro}}{\text{desviación típica de la media muestral}} \sim t_{n-1} \text{ de Student.}$$

El test t concierne a la precisión de la media de una muestra de observaciones, y posibilita poner a prueba la significación de una hipótesis sobre la media poblacional. Si nuestro sedante no tuviese efecto alguno ($\mu = 0$), sería de esperar que la media muestral $\bar{X}$ estuviese en el intervalo:

$$\mu \pm 2{,}262 \cdot \frac{\hat{S}}{\sqrt{n}} = 0 \pm 2{,}262 \cdot \frac{1{,}23}{\sqrt{10}} = (-0{,}88, +0{,}88).$$

Como la media muestral es +1,58, podemos rechazar la hipótesis nula: el nuevo medicamento es efectivo.

Lecturas recomendadas

BELL, E.T., *Los grandes matemáticos*, Buenos Aires, Losada, 2010.

BOYER, C., *Historia de la matemática*, Madrid, Alianza Editorial, 2007.

FISHER, R.A., *Statistical methods, experimental design and scientific inference*, Oxford University Press, 2003.

GRIMA, P., *La certeza absoluta y otras ficciones*, Barcelona, RBA, 2010.

HACKING, I., *La domesticación del azar*, Barcelona, Gedisa, 1995.

HALD, A., *A History of Mathematical Statistics from 1750 to 1930*, Nueva York, Wiley, 1998.

PEÑA, D., *Fundamentos de estadística*, Madrid, Alianza, 2008.

PORTER, T., *The Rise in Statistical Thinking, 1820-1900*, Princeton University Press, 1986.

RIVADULLA, A., *Probabilidad e inferencia científica*, Barcelona, Anthropos, 1991.

STEWART, I., *Historia de las matemáticas*, Barcelona, Crítica, 2008.

STIGLER, S., *The History of Statistics*, Harvard University Press, 1986.

Índice